Readability index of essays as an alternative to the scoring procedure in L2 academic writing

Majid Nemati, Masoud Azizi*

Faculty of Foreign Languages and Literatures, University of Tehran, Iran

*Corresponding author: e-mail address: m.azizi@ut.ac.ir (M Azizi)

ABSTRACT

Samples of participants' writing were scored by two raters using TOEFL writing scoring rubric. The readability index of each text was calculated through the use of six readability formulae and graphs, i.e., Flesch-Kincaid index, Reading Ease index, FOG index, SMOG formula, Fry's graph, and Dale-Chall readability index. The scores given to each essay were later compared to the obtained readability indices through the use of Spearman *rho* correlation coefficient formula. The correlation coefficients obtained ranged from .05 to .15, none of which significant. This indicates that readability index of a text and the writing assessment procedure through holistic rubrics are dealing with two different constructs and have very little in common. This also calls into question the reliability and validity of some computerized assessment programs such as PEG, LSA, or E-rater, which take into account factors very similar to those examined in readability formulae.

Keywords: Readability; Automated essay scoring; Writing

INTRODUCTION

Evaluation has always been affected by numerous factors most of which not of any interest to the stakeholders involved. The literature is full of studies aimed at issuing these factors and attempts to identify and minimize, if not to eradicate, their effects. Among these factors are: students' gender [2,32,36], their ethnic background [27], socioeconomic status [25], behavior [32], and handwriting [29,33,40,41].

As Klein and Taub [29] mention, "the lack of objectivity may often stem from a combination of bias factors, rather than from a single one" (p. 135). One of the best domains in which the subjectivity of human rating reveals itself is the case of second language writing. Writing has found its role in all language instruction courses as well as all language proficiency tests such as IELTS and TOEFL mostly due to its recognition as an important skill and an indication of literacy in a language [15]. However, teaching writing skill cannot be separate from testing it, but the difficulties involved in rating compositions have added to the subjectivity of assessing this very important skill. In other words, scores given to a text should be consistent both when different raters rate the same text and when a rater rates one piece of writing more than once [23]. However, in reality, this proves to be more easily said than done.

Raters are affected by too many factors which cannot be totally controlled or eliminated. They may be affected as much by their own cultural contexts and experiences as by the quality of the written texts. Even when texts are double marked, raters can differ in what they look for in writing and the standards they apply to the same text [46]. Raters' background experience may also obscure their judgments. Research has shown that raters from different disciplines apply different criteria to nonnative English writing samples [7; 42; 45]. Also, raters familiar with students' L1 rhetorical conventions tend to be more accepting of L2 essays showing L1 traces, than other raters [22; 30]. Another factor affecting raters is their rating experience. Keech and McNelly [26] comparing the holistic rating of three rater groups found that students' (group 1) ratings were significantly lower than those of teachers (group 2), and novice teachers' (group 3) ratings were in between.

Moreover, Sweedler-Brown [44] observed that rater trainers were harsher in their assessment of L2 writings than less experienced raters. Cumming [11] reports the same findings in case of L2 and Breland and Jones [4] did so in case of L1.

There are many other factors which may influence how raters, even trained raters, assess a piece of writing. So it seems that any so-called objective method of assessment which still involves human raters is more or less subjective. As a result there have been many attempts to rater-proof the assessment of writing. Using automated assessment programs, such as computerized scoring systems, has been one of such attempts.

Computerized Scoring Systems

Searching the Internet, one can find some programs which offer language learners and involved stake holders the possibility of assessing students' writing samples on-line. As such, different institutes have devised different programs for the purpose of rating their clients' writing samples, each of which considering some features of text as the variables involved in the task of predicting learners' writing ability. Page [1968, as cited in Weigle 46] in his approach to computerized scoring, called Project Essay Grade (PEG), used regression analysis to "determine how well a number of variables such as average sentence length, number of paragraphs, and punctuations could predict the scores given by human raters to a fairly large set of training essays" (p. 234).

More recent studies on PEG [37; 38] have shown that scores given by PEG are of high correlation with scores given by single human rater as well as, and even better than, pairs of raters. However, Chung and O'Neil [9] point out some limitations in the use of PEG. Since PEG does not take into account the meaning and message of a text and only pays attention to surface features of them, it faces problems in considerations of construct validity. Also, a PEG system should be specifically developed for each set of essays used since scores derived from PEG are meaningful only in respect to the set of essays being used. Finally, no exact description of the variables PEG takes into account has ever been published. As a result, very little is known about the relative weight of each variable in determining an essay score.

Latent Semantic Analysis (LSA) is another approach to computer essay scoring, which is "both a computational model of human knowledge representation and a method for extracting semantic similarity of words and passages from text" [18, p. 1]. LSA, unlike PEG which only takes into account the surface features of texts, is based on comparing semantic content of words used in essays. As a result, LSA is more appropriate in assessing writing in content-area courses [46].

LSA, like PEG, is quite reliable. In a study, [18] reported that while the correlation between pairs of human raters was .83, the correlation of LSA scores with scores given by human raters was .80. As Chung and O'Neil [9] point out, LSA as a web-based application can be advantageous to students as it gives them the opportunity to receive immediate feedback on their essays. Moreover, LSA uses both relative and absolute scoring methods; that is, it is possible to compare an essay either to other essays within the same sample, or to an outside source document, e.g. to that of an expert. However, as Weigle [46] mentions, LSA has a disadvantage: it does not take into account the word order making every possible combination of words in a sentence equivalent.

E-rater, developed by the Educational Testing Service (ETS), is a more recent approach used to rate essays written for the Graduate Management Admission Test (GMAT) in conjunction with human raters. It is designed to analyze essays based on the features specified in scoring guides used by human raters. Like PEG, it uses regression analysis of a large number of variables on scores of training essays in order to predict the scores for the rest of the essay set. However, unlike PEG, it takes into account more variables such as syntactic structure, rhetorical structure, and topical analysis [8].

Despite all the developments in computerized essay scoring, there are many who argue against the use of these systems. Most of these programs do not reveal the underlying factors which they take into account when rating the samples. Drechsel [13] states, "not only does this method of assessment disregard decades of research on the writing process, but it also assumes a theory of reading that goes backward in time to New Criticism – when all there was to a page of writing was a page of writing" (p. 384).

But still remains unclear the extent to which computerized scoring systems can replace human raters. Breland [3] believes that "grading is a high-stakes event that can affect other important events, such as college admission; accordingly, grading seems an unlikely task for the computer" (p. 255). He further suggests that computers can be used to help students edit their work and to help teachers examine different features in their students' writing which they may have overlooked otherwise.

Kukich [31] believes that validating automated scoring systems in this way is to some extent circular since the primary objective of these programs is to reproduce the scores given by human raters, while at the same time, expert raters are taught and try to apply a specific scoring rubric as consistently as possible without any "personal or professional feelings about the quality of the writing sample" (p. 17). In other words, raters are trained to copy what automated scoring systems do while these programs try to emulate what human raters are taught to do. Thus, many researchers have questioned the significance of high correspondence between these two systems [1,5,10,24,39,47].

Although those designing such programs do not reveal that much about the factors they take into account, based on the information available it seems that many factors examined by such automated assessment programs are very similar in nature to those considered in readability formulae. The factors examined by readability indices, having accounted for more than 250 variables indentified in a text [28], and having been used in most of such automated programs, should turn out to be highly correlated with scores given to learners' writing samples using writing scoring rubrics.

Readability indices

Readability is defined as "what makes some texts easier to read than others" [14, p. 3]. Wimmer and Dominick [48] defines readability as the "sum total of the entire elements and interactions that affect the success of a piece of printed material" (p. 331). McLaughlin [35, p.188] defines readability as "the degree to which a given class or people find certain reading matter compelling and, necessarily, comprehensible." This definition relates the text with a class of readers of known characteristics such as reading skill, prior knowledge and motivation [14].

Readability is more broadly defined as the "comprehensibility of written text" [21, p. 306]. As a result, readability formulae aim at predicting and quantifying the comprehensibility of a text for its specified readers. The methods of quantifying the readability of texts are very much the same for most formulae. As Stoke [43] explains, a series of graded passages is taken as the criterion and is used to identify variables such as average word length, sentence length, number of polysyllabic words per N sentences, etc. Since no limit to the number of variables can be specified, only those variables which correlate best with the grade level of the passages are combined using a multiple regression analysis.

There are so many readability formulae available for use, each taking into account a number of different but related text variables. Flesch-Kincaid readability index, Flesch's Reading Ease, Dale-Chall readability index, Gunning's Fog index, Fry's readability Graph, and McLaughlin's SMOG readability index are some of the more famous ones.

In his dissertation, Flesch introduced his first readability formula for measuring adult reading material. He used two variables: affixes and personal references such as personal pronouns and names. It soon proved to be very useful. In 1948, he published his second formula, *the Reading Ease formula*, in which he used two variables: the number of syllables and the number of sentences for each 100-word sample. The formula for the updated Flesch Reading Ease score is:

Score = 206.835 – (1.015 x ASL) – (84.6 x ASW)

Where:

Score = position on a scale of 0 (difficult) to 100 (easy), with 30 = very difficult and
70 = suitable for adult audiences.

ASL = average sentence length (the number of words divided by the number of sentences).

ASW = average number of syllables per word (the number of syllables divided by the number of words).

This formula correlates .70 with the 1925 McCall-Crabbs reading tests and .64 with the 1950 version of the same tests [14]. In order to further simplify this formula, Farr, Jenkins, and Paterson [16] modified it as follow:

New Reading Ease score =
1.599 nosw – 1.015 sl – 31.517

Where:
nosw = number of one-syllable words per 100 words;
sl = average sentence length in words

This formula, also called Flesch-Kincaid formula, the Flesch Grade-scale formula as well as the Kincaid formula, correlates better than .90 with the original Flesch Reading Ease Formula.

PURPOSE OF THE STUDY

One major problem in teaching writing is the evaluation of this skill. In addition to being subjective in nature, rating compositions has always been costly and time consuming. Not all institutes involved in writing assessment enjoy having available a number of expert raters. Most often, essays are assessed by those who only happen to teach the writing course and are not experienced enough to be able to rate the essays in a less subjective and more objective fashion. What they do is to compare each student's performance with those of others based on the impression their written work has on them. Moreover, even trained raters are affected by too many factors unrelated to the construct being measured controlling which almost impossible. This violates test fairness according to which students' scores should not be determined by those who happen to be the raters [23]. The existence of such problems has stimulated many attempts on the part of researchers to develop rater-proof scoring procedures. However, these programs can take into account almost nothing but the surface features of a text. Coherence, for example, seems an impossible notion to be captured by such automated assessment programs. As a result, this study sought to find out whether the surface features of a text, accounted for by readability indices, can be a good indicator of second language learners' writing ability assessed by human raters using a holistic writing rubric. Besides, the results of this study could be used as evidence to evaluate computerized assessment programs. They take into account factors very similar to those examined by readability formulae. As such, the presence or absence of the relationship between readability indices and raters' scores could further support the use of such programs or otherwise, could question their usefulness. As such, the following research question was formulated: To what extent readability indices of texts written by learners of English as a foreign language could substitute the scores given to the same texts by human raters?

MEHTODS

Participants

For the purpose of the present study, 16 male and 38 female Iranian upper-intermediate learners studying English in TOEFL and IELTS courses in an English language institute in Tehran, Iran, had participation. All these students had already been tested using mock TOEFL (pbt) and mock IELTS with those scoring 5 or higher in IELTS and 61 or higher in TOEFL (iBT) being placed in IELTS and TOEFL courses respectively (It should be noted that the applicants take the TOEFL pbt exam as a placement test and their scores are adapted to TOEFL iBT scoring guide). The participants' age ranged from 20 to 28 and they had all received writing instruction as a part of their instruction program in that institute.

Data Collection

To gather learners' writing samples, a topic was chosen from the TOEFL Writing Topic Booklet available on ETS homepage. The chosen topic required learners to create an argumentative piece of writing. The participants were given 40 minutes for planning and writing about the given topic. Although TOEFL gives test takers 30 minutes for fulfilling its writing task, it was decided that similar to Task 2 on ILETS participants be allowed to complete the writing task in 40 minutes. This decision was made based on the piloting done before the study began, and it was due to the fact that time pressure could affect participants' performance, and the results of the study. The samples were collected at the end of their writing course when the students had already received the instructions for writing skill. The writing test was administered as a part of participants' instruction and during their class hours. In order to avoid the Hawthorn and halo effects, measures were taken not to clue students to the fact that they are participating in a research project.

In order to avoid the effects of handwriting on the raters, all the gathered samples were typed by the researchers. The researchers were cautious to type them as they were actually written by the participants, that is, all misspellings, wrong punctuations, and other types of mistakes were typed exactly as they had appeared in the scripts. The samples were then given to two experienced raters to be rated based on TOEFL holistic writing rubric. Before rating the samples, a meeting was arranged with both raters and the procedure and the type of scoring guide were explained to them. However, they were not clued in on the purpose of the study.

To make the ratings more precise and in order to have a better range of scores, the raters were required to make one more decision for each sample. Based on the scoring rubric, each writing was assigned to a level ranging from 0 to 6, with levels 0, 1, and 2 almost never occurring in case of learners at upper-intermediate level. This could severely limit the range of scores and as a result understimate the correlation between the readability indices and the scores given by the raters. Therefore, each level from 1 to 6 was further divided into 3 sub-levels. For example, level 3 was divided into 3- (read as 'three minus'), 3, and 3+ ('three plus'). If a sample were not good enough to be assigned to level 4, but at the same time a score of 3 could not be justified for it, that writing would be assigned to 3+. On the other hand, when a sample was not good enough to be assigned to level 3, but it was not that much bad to be relegated to level 2, that piece of writing would be assigned to level 3-. Then, 1- was entered to SPSS as 0, 1 as 1, 1+ as 2, 2- as 3, etc. As such, the possible range of scores for samples was 1 to 17.

According to Brown, Glasswell, and Harland [6], when the raters are trained to judge based on a scale rubric, the consensus estimates including the percent exact agreement and adjacent agreement between raters could act as the best measures of agreement. The percent exact agreement obtained was 67% and the adjacent agreement index was 91%. A correlation coefficient of .93 was obtained in the case of inter-rater reliability. Also after having raters re-rate 25% of the randomly selected samples, the researchers calculated the correlation coefficient between the scores given by the raters in their first and second attempts. The obtained correlation coefficient for rater 1 was .97 and for rater 2 was .90. Table 1 gives the descriptive statistics of the scores given by the raters to the samples written by the participants.

Table 1 .Descriptive Statistics for the Mean Scores Given by Raters

	N	Minimum	Maximum	Mean	Std. Deviation
Rater	54	6	15	9.61	2.351

The typed samples were also analyzed in terms of their readability indices through the use of Flesch-Kincaid readability index, Flesch's Reading Ease, Gunning's Fog index, McLaughlin's SMOG readability index, Fry's readability Graph, and Dale-Chall readability index. In order to calculate the readability indices of the texts, the researchers used some of the websites on the Internet such as http://www.online-utility.org/english/readability_test_and_improve.jsp, http://www.harrymclaughlin.com/SMOG.htm, http://www.educationalpsychologist.co.uk/fry_readabilityprogram.htm, http://www.interventioncentral.org, which offered some computerized programs to assess the readability of the texts. However, most readability formulae were considered and analyzed by using more than one program. Although the obtained scores were different from each other, they highly correlated with each other. The correlation coefficients obtained between each pair of readability indices calculated by two different programs ranged from .86 to .95.

RESULTS

The obtained readability indices of the writing samples and the scores given by the raters using the holistic scoring rubric of TOEFL were compared with each other through using Spearman *rho* correlation coefficient provided by SPSS. Table 2 summarizes the correlation coefficients obtained between the six readability indices calculated for learners' writing samples and the scores given by human raters.

Table 2. The Summary of the Correlation Coefficients Obtained

Indices	Raters' scores	sig.
Flesch-Kincaid	.08	.56
Flesch's Reading Ease	-.15	.27
Gunning's Fog index	.05	.69
SMOG	.07	.59
Fry's Graph	-.05	.71
Dale-Chall index	.05	.70

As evident in the above table, the six readability formulae appeared to be of almost no relationship with the scores given to learners' writing samples by human raters using a holistic scoring guide. The correlation coefficient obtained ranged from .05 to .15, which was not statistically significant for any of the indices (p > .01).

DISCUSSION

Based on the figures obtained, it is obvious that both readability formulae and human raters are dealing with two completely different constructs. In order to see how much an index has in common with another index or how much variation in one construct is accounted for by another construct, or even to see if two different measures examine the same construct, the coefficient of determination should be obtained, that is, the correlation coefficient index should be squared. The result will be the amount of the common variance.

The largest amount of correlation coefficient observed among the examined readability formulae was .15, the square of which equals .0225 which is ignorable. So it is plausible to conclude that the readability formulae and the raters scoring based on the holistic scoring rubric measure two different constructs and have nothing in common.

Readability indices take into account different surface text features such as average sentence length, the number of syllables per 100 words, percentage of words out of some especial word lists such as that of Dale-Chall word list, percentage of polysyllabic and monosyllabic words per 100 words, the number of personal pronouns and proper nouns or prepositions, average number of syllables per word, and the number of affixes in a text. These factors, although directly related to a text, have turned out to be of no relationship with the assessment of language learners' writing ability. This could be either due to the fact that these two measures assess two different constructs or for the problems readability indices face with.

The absence of common variance between what readability indices measure and what holistic scoring rubrics measure in writing assessment warns us about the use of computerized programs used to assess learners and test takers' writing samples. Programs such as Project Essay Grade (PEG), Latent Semantic Analysis (LSA), and E-rater developed by ETS take into account surface text features very similar to those examined by readability indices. For example, PEG uses sentence length, number of paragraphs and punctuations in order to predict the scores raters will give to writing samples. Also, PEG does not take into account the meaning and message of a text [9] and LSA, like readability formulae, does not consider the word order [46]. While there are studies confirming the validity of such programs, the present study calls for more precise examination of the extent to which one can rely on such assessing programs in her decision makings.

Therefore, there is a lack of clarity as to the extent to which computerized scoring systems and the factors and text features they examine can replace human raters in the task of writing assessment. It seems that human rater is an indispensable element of writing assessment even in large scale evaluations. As Breland [3] states, "grading is a high stake event that can affect other important events, such as college admission; accordingly, grading seems an unlikely task for the computer" (p. 255).

Teachers who tend to assess their students' writing samples using the available writing assessment programs on the Web without knowing or paying attention to the criteria and factors these programs take into consideration in the task of assessment are the first group of people who should be careful. The present study shows that most surface text features used in readability formulae, which are also used in most of these programs, are of no relationship with writing abilities of language learners. As a result, language teachers and assessors should be careful about the options they have in writing assessment. Moreover, the findings of this study warn those involved in the task of assessment to be more cautious about the factors and features they take into account while assessing learners' writing samples, especially if they plan to design a computerized program to do the task of assessment. Such factors could not be good predictors of learners' writing ability.

Researchers, teachers, and other users should doubt the reliability of programs such as PEG, LSA, E-rater and similar programs available to them on the Internet or exclusively used by some particular organizations. They should approach such programs with extreme care and always have a second thought before making any decision about replacing them for human raters. More studies need to be conducted in order to make sure that the assessment of learners' writing ability is fair and is not affected by factors unrelated to the construct being examined.

It seems clear that such programs cannot also be useful to second language learners for the purpose of self assessment. Having considered the surface features which are of no relationship with learners' writing ability, and having failed to take into account more important factors such as coherence in the text, such programs appear to be of little help to language learners who seek for feedback and evaluation of their writing samples. Learners should also be informed that the factors these programs may take into account could be different from those considered by human raters.

Moreover, it is a common belief among learners that the more low frequency words they use in their writing, the more they can impress the rater, and the higher their scores would be. Lack of any relationship between raters' scores and the readability formulae which take into account low frequency words shows that there is not sufficient evidence to confirm this belief, and there are other factors which should be pursued if they are willing to improve their writing abilities.

REFERENCES

1. Bennett, R.E., & Bejar, I.I. Validity and automated scoring: It's not only the scoring. Educational Measurement: Issues and Practice.1998; 17**,** 9–17.

2. Bolger, N., & Kellaghan, T. Method of measurement and gender differences in scholastic achievement. Journal of Educational Measurement. 1990; 27, 165–174.

3. Breland, H. M. Computer-assisted writing assessment: The politics of science versus the humanities. In: White E.M., Lutz W.D., & Kamssikiri S, editors. Assessment of writing: politics, policies, practices. New York: Modern Language Association of America; 1996. P. 249-256.

4. Breland, H. M., & Jones, R. J. Perception of writing skills. Written Communication. 1984; 1(1), 101-119.

5. Brent, E., & Townsend, M. Automated essay grading in the sociology classroom. In: Ericsson P.F. & Haswell R, editors. Machine scoring of student essays: Truth and consequences. Utah State University Press, Logan, UT; 2006. P. 177-198.

6. Brown, G.T.L., Glasswell, K., & Harland, D. Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. Assessing Writing. 2004; 9, 105–121.

7. Brown, J. D. Do English and ESL faculties rate writing samples differently? TESOL Quarterly. 1991; 25, 587-603.

8. Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. Enriching automated essay scoring using discourse marking. Princeton, NJ: Educational Testing Service; 2001.

9. Chung, G.K.W.K, & O'Neil, H. F. Jr. Methodological approaches to online scoring of essays. Report No. CSE-TR-461, 1997. ERIC Document Reproduction Service No. ED 418101.

10. Clauser, B.E., Kane, M.T., & Swanson, D.B. Validity issues for performance-based tests scored with computer-automated scoring systems. Applied Measurement in Education. 2002. 15(4), 413–432.

11. Cumming, A. Expertise in evaluating second language composition. *Language Testing. 1990; 7*, 31-51.

12. Dale, E. & Chall, J. S. The concept of readability. Elementary English. 1949; 26, 19-26.
13. Drechsel, J. Writing into silence: Losing voice with writing assessment technology. Teaching English in the Two-Year College. 1999; 26(4), 380-387.
14. DuBay, W.H. The principle of readability, National Adult Literacy Database, Inc, 2004. Retrieved October 16, 2007, from http://www.nald.ca/fulltext/readab.pdf
15. Ediger, A. Teaching children literacy skills in a second language. In: Cece-Marcia M, editor. Teaching English as a second or Foreign Language. 3rd ed. U.S. Heinle & Heinle. 2001. P. 153-164.
16. Farr, J. N., Jenkins, J. J., & G. Paterson, D. Simplification of the Flesch Reading Ease Formula. Journal of Applied Psychology. 1951; 35(5), 333-357.
17. Flesch, R. A new readability yardstick. Journal of Applied Psychology. 1948; 32, 221-233.
18. Foltz, P. W., Laham, D., & Landauer, T.K. Automated essay scoring: Applications to educational technology. Proceedings of EdMedia, 1999. Retrieved September 20, 2007, from http://www.psych.nmsu.edu/~pfoltz/reprints/Edmedia99.html
19. Fry, E. B. The readability graph validated at primary levels. The Reading Teacher. 1969; 22, 534-538.
20. Gunning, R. The technique of clear writing. New York: McGraw-Hill; 1952.
21. Hewitt, M., & Homan, S. Readability. In: Thompson R.A, editor. Classroom reading instruction Dubuque, IA: Kendal/Hunt; 1991. P. 305-318.
22. Hinkel, E. Native and nonnative speakers' pragmatic interpretations of English texts. TESOL Quarterly. 1994; 28, 353-376.
23. Huot, B. Reliability, validity, and holistic scoring: What we know and what we need to know. College Composition and Communication. 1990; 41, 201–213.
24. James, C.L. Validating a computerized scoring system for assessing writing and placing students in composition courses. Assessing Writing. 2007; 3, 167-178.
25. Jussim, L., & Eccles, J. S. Teacher expectations: Construction and reflection of student achievement. Journal of Personality and Social Psychology. 1992; 63, 947–961
26. Keech, C. L. & McNelly, M.E. Comparison and analysis of rate responses to the anchor papers in the writing prompt variation study. In: J.R. Gary and L.P. Ruth (Eds.), Properties of writing tasks: A study of alternative procedures for holistic writing assessment. Berkeley: University of California, Graduate Scholl of Education, Bay Area Writing project; 1982.
27. Keith, T. Z., & Reimers, T. M. Parental involvement, homework and T.V. time: Direct and indirect effects on high school achievement.Journal of Educational Psychology. 1986; 78, 373–380.
28. Klare, G. R. Readability. In: P. D. Pearson (Ed.), Handbook of reading research, New York: Longman; 1984. P. 681-744.
29. Klein, J. & Taub, D. The effect of variation in handwriting and print on evaluation of student essays. Writing Assessment. 2005; 10, 134-148.
30. Kobayashi, H. & Rinnert, C. Factors affecting composition evaluation in an EFL context: Cultural rhetorical pattern and readers' background. Language Learning. 1999; 46, (3), 397-437.
31. Kukich, K. Beyond Automated Essay Scoring. IEEE Intelligent Systems. 2000; 15(5), 22–27.
32. Manke, M. P., & Loyd, B. H. An investigation of non achievement-related factors influencing teachers' grading practices. Paper presented at: The Annual meeting of the National Council on Measurement in Education; 1990, April; Boston.
33. Marshall, J. C., & Powers, J. H. Writing neatness, composition errors and essay grades. Journal of Educational Measurement. 1969; 6, 97–101.
34. McCall, W. A., & Crabbs, L. M. Standard test lessons in reading. New York: Teachers College, Columbia University Press; 1925, 1950, 1961, 1979.
35. McLaughlin, G. H. Proposals for British readability measures. In: J. Downing J & Brown A.L, editors. The Third International Reading Symposium. London: Cassell; 1698, p. 186-205.
36. Natriello, G., & McDill, E. L. Performance standards, student effort on homework, and academic achievement. Sociology of Education. 1986; 59, 18–31.

37. Page, E.B. Computer grading of student prose, using modern concepts and software. Journal of Experimental Education. 1994; 62 (2), 127–142.
38. Page, E.B., & Peterson, N.S. The computer moves into essay grading: Upgrading the ancient test. Phi Delta Kappan. 1995; 76 (7), 561–569.
39. Powers, D.E., Burstein, J.C., Chodorow, M.S., Fowles, M.E., & Kukich, K. Comparing the validity of automated and human scoring of essays. Journal of Educational Computing Research. 2002; 26 (4), 407–425.
40. Russell, M. The influence of computer print on rater scores. Technology and Assessment Study Collaborative. CSTEEP, Boston College; 2002.
41. Russell, M. & Plati, T. Mode of Administration Effects on MCAS Composition Performance for Grades Four, Eight and Ten. A report submitted to the Massachusetts Department of Education by the National Board on Educational Testing and Public Policy, 2000.
42. Santos, T. Professors' reactions to the academic writing of nonnative-speaking students. TESOL Quarterly. 1988; 22 (1), 69-90.
43. Stokes, A. The reliability of readability formulae. Journal of Research in Reading. 1978; 1, 21-34.
44. Sweedler-Brown, C.O. The influence of training and experience on holistic essay evaluation. English Journal. 1985; 74 (5), 49-55.
45. Sweedler-Brown, C.O. ESL essay evaluation: The influence of sentence-level and rhetorical features. Journal of Second Language Writing. 1993; 2(1), 3_17.
46. Weigle, S. Assessing writing. Cambridge: Cambridge University Press; 2002.
47. Williamson, D.M., Bejar, I.I., & Hone, A.S. 'Mental Model' comparison of automated and human scoring, Journal of Educational Measurement. 1999; 36 (2), pp. 158–184.
48. Wimmer, R.D., & Dominick, J.R. Mass Media Research: An Introduction. 8th ed. Australia: Thomson Wadsworth Publishers, 2005.