THE ANALYSIS ON DIFFERENCE BETWEEN MACHINE SCORING AND HUMAN SCORING IN WRITING ASSESSMENT

Fu Lei
National University of Uzbekistan Named after Mirzo Ulugbek, Tashkent, Inner Mongolia University of Finance and Economics, Hohhot, China

Annotation. The research chooses 20 writing samples from the automated scoring system and compares the results of machine scoring and human scoring. Based on the analysis on the differences and the reasons leading to the differences, the research concludes the results of the automated evaluation system are reliable, and human assessment should be added to high-risk tests to ensure accuracy. According to the analysis on text features, the research provides some suggestion for language teaching and learning.

Key Words: Scoring, Difference, Analysis, Assessment

YOZMA ISHNI TEKSHIRISHDA DASTUR VA SHAXS TOMONIDAN BERILGAN BAHO O‘RTASIDAGI FARQNI TAHLIL QILISH

Fu Lei
irzo Ulug’bek nomidagi O’zbekiston Milliy unversiteti mustaqil tadqiqotchisi M

Annotatsiya. Tadqiqot avtomatlashtirilgan ball tizimidan 20 ta yozma maqola tanlab olindi va ularni inson bahosi bilan solishtirilindi. Farqlarga olib keladigan sabablarni tahlil qilish asosida tadqiqot avtomatlashtirilgan ball tizimi natijalari ishonchliligi aniqlanib, ammo aniqlikni ta‘minlash uchun yuqori xavfli testlarga inson bahosini qoshish kerak degan xulosaga kelindi. Matn xususiyatlarini tahlil qilish jarayonida tadqiqot tilni o‘qitish va o‘rganish bo‘yicha ba‘zi takliflar taqdim etildi.

Kalit so‘zlar: reyting, farq, tahlil, baholash.

АНАЛИЗ РАЗНИЦЫ МЕЖДУ ОЦЕНКОЙ, ВЫСТАВЛЕННОЙ ПРОГРАММОЙ И ЧЕЛОВЕКОМ ПРИ ПРОВЕРКЕ ПИСЬМЕННЫХ РАБОТ

Фу Лей
независимый соискатель Национальный университет Узбекистана имени Мирзо Улугбека

Аннотация. В ходе исследования были отобраны 20 письменных работ из автоматизированной системы оценки, результаты которых были сравнены с их оценкой человеком. На основании анализа различий и причин, приводящих к различиям, исследование пришло к выводу, что результаты автоматизированной системы оценки надежны, но для обеспечения точности к тестам высокого риска следует добавить оценку, проводимую человеком. На основе анализа текстовых особенностей, исследование дает некоторые предложения по преподаванию и изучению языка.

Ключевые слова: рейтинг, разница, анализ, оценка.

Introduction. Writing plays an important part in language learning process and presents language learners’ ability. Language learners can improve their writing ability based on immediate feedback. In recent ears, Automated Writing Evaluation System (AWE system) is applied widely in the college writing courses. Language learners can practice writing and correct the mistakes based on the suggestion given by the system, so that they can improve the writing skills and autonomous learning. Meantime, the system can reduce the great burden teachers must experience when they do paper-pencil correcting.

The research is to answer three questions:

i. Are the results of machine scoring consistent with the results by human scoring?

ii. If there are some differences, what are the reasons leading to them?

iii. What are the text features of high-score writing samples?

The research chooses 20 samples from iWrite 2.0, compares the scores provided by machine and human raters, and analyzes the reasons leading to the differences. The research also provides some suggestion on writing teaching and learning based on the text features by Eng-Editor.

A Literature Survey

The first AWE, Project Essay Grade (PEG), was invented by Ellis Page. He aimed to reduce the burden of teachers and hypothesized that the system can provide the score of the composition based on some text features. Besides PEG, there are some other automated evaluation systems, such as Intelligent Essay Assessor (IEA), E-rater, IntelliMetric, Criterion and so on. (Dikli, 2006)

AWE have been applied in many high-stake tests. Different systems analyze the text from different indicators, and most of them may include vocabulary, syntax, text, mistakes and so on. (□&□□2018) Take TOEFL as an example. Vocabulary is assessed from complexity and average length; structure is from organization and development; mistakes include grammar mistakes, genre, usage and so on; content is assessed from key words and collocation. (Deane, 2013) In order to ensure the accuracy of test, the writing products must be assessed by both the system and human in the high-stake tests to ensure the accuracy.

In China, some famous automated evaluation systems have been invented, such as www.Pigai.org (Pigai system), iWrite 2.0 system. The writer searches for the articles on this topic and finds some researchers have conducted some studies on the evaluaiton system, including quantitative studies and qualitative studies. Pigai system provides the score based on the following indicators: average length of sentence, usage of different sentence structure, spelling, grammar, and vocabulary. However, the system provides less advice on structure, content, and organization. (□, 2013)

Some quantitative studies focus on the prediction of scores according to text features. Writing sections in College English Test 4/6 (CET4/6) are assessed by human from content, organization, and mistakes, and scored in different levels. Some researchers set a model with more than 10 indicators and find that the automated evaluation system depends on the quantitative features. (□&□□2018)

The automated teaching and evaluation system iWrite 2.0 can immediately evaluate students' compositions from four dimensions (language, content, organization, and mechanics) and highlight their errors and error types. □Li, & Xiao□2020) The system is the first one to build a word association network with less than 5 key words to assess the production from the relevance and coherence of the essay. It also represents the writing assessment for the language teachers to help improve teaching and learning.

The automated evaluation system can reduce the burden of teachers and help language learners improve writing products. However, it could not assess the organization, relevance, and argument. If it needs a more accurate assessment, the products should be assessment by human.

Research Methodology

The research is to analyze the reasons leading to the difference between the machine scores and human scores based on the text features.

Writing samples and writing task

The research chooses 20 samples from a writing competition by 20 students of from different majors, such as English language, Accounting, International business, Business administration and so on. The writers come from different grades, including 6 freshmen, 7 sophomores, 5 juniors, 1 senior and 1 first-grade post-graduate. More than 80% are major in humanity and management. All of writing samples are divided into 3 grades based on the machine scores, including 5 from Grade 1,10 from Grade 2, and 10 from Grade 3.

The writing task is to read a paragraph about the contradictory views on silent carriages, then complete a composition on iWrite 2.0 system about this issue stating opinions and explaining reasons. The essays can be scored by the system.

Machine scoring and human scoring

After completing the task, all the essays can be scored by the system, and by three raters. The three raters are experienced language teachers, and one of them has experiences of assessing writing part in College English Test 4 (CET 4). There are some differences among the scores by different raters, and some differences between the machine scores and the average scores of human scorings. (Table 1)

Sample	**Rater 1**	**Rater 2**	**Rater 3**	**Average score**	**Machine scoring**	**Difference**
1	41.5	47.5	46.0	45.00	46.0	1.00
2	41.5	44.0	43.5	43.00	43.6	0.60
3	37.5	40.0	43.0	40.17	42.2	2.03
4	34.0	40.0	42.5	38.83	38.3	-0.53
5	34.0	41.5	45.0	40.17	38.7	-1.47
6	36.0	40.0	40.0	38.67	40.2	1.53
7	32.5	41.5	44.5	39.50	40.2	0.70
8	34.0	41.0	42.0	39.00	38.8	0.20
9	37.5	39.5	44.0	40.33	36.5	-3.83
10	36.0	39.0	38.0	37.67	37.0	-0.67
11	20.0	25.0	38.5	27.83	39.9	12.07
12	38.0	36.5	42.5	39.00	38.7	-0.30
13	35.0	40.0	44.5	39.83	38.6	-1.23
14	38.5	37.5	42.5	39.50	34.2	-5.30
15	31.5	39.0	39.0	36.50	35.0	-1.50
16	27.5	37.5	42.0	35.67	35.5	-0.17
17	29.0	36.0	41.5	35.50	34.3	-1.20
18	30.5	40.0	42.0	37.50	32.5	-5.00
19	29.0	39.0	41.5	36.50	31.1	-5.40
20	31.5	37.5	42.5	37.17	34.7	-2.47
Average	33.75	39.1	42.25	38.37	43.6	5.23

Table 1. Human scoring and Machine scoring

3.1. Results and Findings

3.2. Results of comparison

The formula for calculating difference between machine scoring and human scoring is to subtract the average score of human scoring from the machine score and count the proportion of samples with score difference of 0-3. In some international tests, the total score is 6, and the samples with a difference of 1 can be in different levels. (白&王，2018) The total score of the writing task is 50 points, and 3 is strict. (Table 2.)

	Difference (point)						
	<0.5	0.5--1	1—1.5	1.5--2	2—2.5	2.5--3	>3
Human scoring vs. Machine scoring	15%	20%	25%	5%	10%	0	25%

Table 2. Human scoring vs. Machine scoring

If the difference between the results of machine scoring and the result of human scoring is less than 3 points, and 75% of the machine scores are consistent with the human scores. Among the 5 samples with the difference of more than 3 points, there are 3 samples with a difference of more than 5 points, and 1 sample with more than 12 points. Consistency between the automated evaluation system and human scoring should be essentially 75% - 80%. (Burstein J & Chodorow M, 1999)

The research is to analyze the results from three dimensions: the reasons leading to differences in human scoring, the reasons leading to great differences of special samples, and the text features of high-score samples.

4.1.1 The reasons leading to differences in human scoring

The three raters are trained on scoring standards before scoring to assure the accuracy of scoring. (Table 3)

Argumentative Writing	
Content/Ideas (40%)	1. Writing effectively addresses the topic and the task; 2. Writing presents an insightful position on the issue; 3. The position is strongly and substantially supported or argued.
Organization/ Development (30%)	1. Writing is well-organized and well-developed, using appropriate rhetorical devices (e.g. exemplification, classification, analysis, comparison/contrast, etc.) to support thesis or to illustrate ideas; 2. Writing displays coherence, progression, consistency and unity; 3. Textual elements are well-connected through explicit logical and/or linguistic transitions.
Language (30%)	1. Spelling is accurate; 2. Writing displays consistent facility in use of language; 3. Writing demonstrates appropriate register, syntactic variety, and effective use of vocabulary.

Table 3 Rating Standards of Human Scoring (https://uchallenge.unipus.cn/c/2022-05-13/512330.shtml)

There are some differences among the scores given by the three raters.

Rater 1 gives the lowest scores, and the scores given by Rater 3 is most consistent with the machine scores. After scoring, the raters are interviewed, and explain the procedure of scoring. Rater 1 gives the scores separately and calculate the score for each sample. Rater 2 has some experiences on the assessment of writing of CET 4. The scoring is influenced by the standards of CET 4 and the daily writing assessment, so that Rater 2 gives highest scores. Rater 3 is the most experienced language teacher and the average score given by her is most consistent with the score by machine.

Although there are some scoring standards, different raters can be affected by other factors. Raters pay more attention to the complexity of sentences, content, grammar, spelling, and if there are more mistakes of spelling and more short sentences, the rater may give lower scores.

4.1.2 The analysis on 5 samples with great differences

There are 5 samples with a difference of more than 3 points, and they are No.9, No.11, No.14, No.18, No.19. The research analyzes the reasons leading to the differences.

No.9 is a typical argumentative. The structure is clear, including personal view, reasons, special cases and conclusion. There are 17 sentences and most of them are combination of short sentences. And the total number of words is less than 500, and the level of vocabulary is below level 5, which is the one college students should get. The content of the text is not clear, and irrelated to the topic, which may be the reason of lower score given by machine and other two raters.

No.11 is a special one. Rater 1 and Rater 2 give the lowest scores because of its deviation from the topic. There are 5 paragraphs, and the structure is lack of relevance. However, the machine scoring system gives a higher score. The reason leading to the differences may be the content is part of reading passage, which is considered as the consistent with the topic.

No.14 consists of 7 paragraphs and 32 short sentences, but the content is clear, and there is an example to support the opinions. Three raters give the higher score compared with the scored by machine. However, there are some grammar mistakes.

No.18 and No.19 have some similarities. The length of writing is not matched with the order of 500 words, and there are so many grammar mistakes which are the reasons of the lowest scores by machine. The raters give the higher scores than system because the organization is more reasonable.

4.2 The analysis on the text features

The automated evaluation system iWrite 2.0 builds a word association network with less than 5 key words and assesses the writing from language, content, organization, and mechanics. Based on the analysis on the texts of 20 samples by Eng-Editor, the research concludes the text features of high-score samples from content, structure, and language.

The content of high-score writing is clear about the writer’s opinions. Besides some common views of the public, there are some personal opinions on the topic and some examples, stories or data to support the opinions.

The structure of most samples is a typical argumentative one. The structure consists of 3---5 paragraphs, including Generation (phenomenon/introduction, personal experience), Body (reasons, different views, benefits), and Conclusion (personal view, benefits, measures). Therefore, the most samples may be supplied with the similar score except some special ones.

As to the language, there are some features based on the analysis by Eng-Editor. (Table 4)

Sample	**VC/Grade**	**SC/Grade**	**LD/Grade**	**Level**
1	6.96	6.32	6.39	6
2	7	7	6.98	6
3	5.91	6.38	5.92	5
4	7.08	6.75	7.25	7
5	4.39	7.37	5.05	5
6	4.88	5.25	4.99	4
7	4.94	3.78	4.95	4
8	4.81	4.04	4.93	4
9	4.98	3.68	4.41	4
10	4.30	6.55	4.78	4
11	0	0	0	0
12	4.07	5.74	4.31	4
13	4.00	3.93	4.06	4
14	4.78	6.97	4.99	4
15	4.31	7.00	4.17	4
16	3.91	3.29	3.87	3
17	4.19	4.07	4.33	4
18	3.99	5.99	3.98	3
19	4.89	7.00	5.02	5
20	3.97	6.32	4.00	4

Table 4 Text Complexity Analysis

(Data coming from languagedata.net/tester)

19 samples are analyzed by Eng-Editor except one, NO.11.The vocabulary classification (VC) is from 3.91 to 7.08, and the syntactic classification (SC) is from 3.29 to 7.37. The difficulty level (LD) is from 3.87 to 7.25. Level 5 is the common level college students should be on according to China's Standards of English Language Ability (CSE), but 68.42% of the samples are below level 5.

The high-score samples consist of more long sentences and compound sentences, and there are more conjunctive words or phrases in the text to connect each sentence. Non-finite verbs are used in writing to achieve a high score, but there are a lot of mistakes which leads to score deduction.

Conclusion

Based on the analysis of differences between human scoring and machine scoring, it is concluded that two different scoring methods are consistent in scoring. Machine scoring can reduce the burden of assessing paper writing of language teachers. Language learners can receive the feedback in time to correct their writing. However, machine scoring can make mistakes. Although human scoring can be affected by the evaluation literacy, experiences of raters and other scoring standards, human scoring should be additional to ensure the accuracy of assessment, especially in some high-stake tests.

Language teachers should help language learner do more writing exercises and practice using more compound sentences and non-infinite verbs in the writing. Language learners also need to increase reading and listening to accumulate vocabulary instead of the repetition of simple words used in writing.

The number of writing samples is limited and instead of a part of language learners in one university, so there are some limitations. To examine the result of improvement, the researchers should do some further studies on follow-up teaching.

References:

[1] Burstein J & Chodorow M. Automated essay scoring for nonnative English speakers [A]. Proceedings of the ACL99 Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing [C]. College Park, MD. 1999. 68-75.

[2] Deane P. On the relation between automated essay scoring and modern views of the writing construct [J]. Assessing Writing, 2013, 18(1): 7-24.

[3] Dikli S. An overview of automated scoring of essays [J]. Journal of Technology, Learning, and Assessment,2006, 5(1): 1-35.

[4] Jin, T., Lu, X., Guo, K., Li, B., Liu, F., Deng, Y., Wu, J., & Chen, G. (2021). Eng-Editor: An online English text evaluation and adaptation system. Guangzhou: LanguageData (languagedata.net/tester).

[5] Li, F. P., & Xiao, L. Y. (2020). A Study on the Teaching of Professional-Oriented English Writing in Applied-Type University Based on I Write 2.0. Creative Education, 11, 1720-1729.

https://doi.org/10.4236/ce.2020.119125

[6]白丽芳,王建.人工和机器评分差异比较及成因分析.外语测试与教学[J].外语测试与教学,2018（3）:44-54.

[7] 何旭良. 句酷批改网英语作文评分的信度和效度研究[J]. 现代教育技术,2013,(5):64-67.

[8]王建，张藤耀. 二语写作文本量化指标与机评分数的关系研究. 外语测试与教学[J].外语测试与教学,2020（3）:12-20.

[9]中华人民共和国教育部.中国英语能力等级量表[EB/OL].(2018-07-25).

http://www.moe.gov.cn/s78/A19/yxs_left/moe_810/s230/201807/t20180725_343689.html.

[10] https://uchallenge.unipus.cn/c/2022-05-13/512330.shtml