Investigating Accuracy of Self-assessment of English Speaking Proficiency Levels by Engineering Students Based on Correlation and Sentiment Analysis

Linyi Qi[a], Xingang Liu[b*], Jiangqin Zhu[c], Jing Wen[d]

University of Electronic Science and Technology of China, Chengdu, China

[a]qilinyi@uestc.edu.cn, [b*]hanksliu@uestc.edu.cn
[c]peggyzjq@uestc.edu.cn, [d]joanniewen@uestc.edu.cn

Abstract. The study investigated the accuracy of self-assessment (SA) and possible reasons for inaccuracies using correlation analysis and sentiment analysis based on machine learning. Results show that there was a moderate correlation between self-assessment and teacher assessment, and a general trend of underestimation in student self-assessment, particularly for high-performing and medium-performing students. Results also show that students' lack of confidence in English language learning, confusion about their own speaking abilities, and lack of a clear understanding of the rating scale are likely to be the reasons for underestimation. Though the accuracy of self-assessment was not validated in this study, it could be argued that SA can still be used as a formative assessment tool to trigger active learning and reflection on own abilities and improve assessment literacy and academic performance through using relevant rating scales or standards.

Keywords: self-assessment (SA); evaluation; correlation; sentiment analysis; underestimation.

1 Introduction

Teacher's evaluation has always been the dominant way of measuring learners' abilities in education. In recent years, however, with pedagogy shifting to student-centered approach and emphasizing students' engagement in almost every aspect of education, an increasing number of studies and educators have started to shift research focus on learners' self-assessment (SA) and explore its implications on targeted teaching and autonomous learning strategies. The practice of self-assessment employing students' reflection of their own abilities has particular implication for engineering students who are a focus of the current study, because the ability to reflect is an IET (the Institution of Engineering and Technology) accreditation criterion when accrediting a programme and also an important graduate attribute which is closely associated with self-regulated and lifelong learning.

M. Yu et al. (eds.), *Proceedings of the 2024 5th International Conference on Big Data and Informatization Education (ICBDIE 2024)*, Advances in Intelligent Systems Research 182,
https://doi.org/10.2991/978-94-6463-417-4_8

2 Related Studies on Self-assessment

2.1 Definition and Value of Self-assessment

Initially SA was regarded a tool of measurement for students to measure their own work against certain standards or criteria. Then the definition was refined as a reflective practice, where students were expected to reflect on not only their own work, but also the process of their learning and development [1]. Under the influence constructivist theories, SA was viewed as a learning tool, and became closely linked with self-regulated learning with emphasis on students' ability to actively monitor, regulate, and adjust their own learning processes [2]. Studies in the 21st century defined SA as an integral component within formative assessment as an ongoing process involving self-generated feedback to inform and improve learning [3]. SA now is broadly understood as a meta cognitive strategy, where learners reflect on their cognition and regulate their learning [4].

Despite its various definitions, SA has been found of great value to students' learning and development. As SA encourages students to reflect on their learning process and strategies, which may result in deeper learning and improvement in their academic performance [5-6]. SA also enables students to be more aware of their strengths and weaknesses, and this in turn tends to increase their confidence and self-efficacy [7]. By engaging in SA, students can foster a sense of ownership of and responsibility for their learning outcomes, and thus become more intrinsically motivated and engaged in their learning [8]. Meta-cognitively, SA helps students to enhance their awareness of their learning processes, and develop their ability to monitor and regulate their cognitive activities [9-10]. In the long term, SA also cultivates students' lifelong learning skills, such as the ability to continuously evaluate one's own performance and to identify areas for further improvement [11-12].

2.2 Studies on Consistency between Student Self-assessment and Teacher Assessment

A good number of studies have focused on the correlation between SA and teacher assessment (TA). Methods such as quantitative studies, qualitative studies, mixed-methods studies, empirical studies, longitudinal studies or case studies have been used. The research findings vary significantly. For example, the study of Lindblom-ylanne, et all showed a significantly positive correlation between SA and TA [12]. The results of Karnilowic's research indicated that, relative to teacher evaluation, low-achieving students were less accurate than high-achieving students; higher-achieving students tended to underestimate themselves while lower-achieving students tended to overestimate themselves [13]. The same conclusion about accuracy and tendency of SA among high- and low- performing students was also drawn in the study of Zvacek et al, but they also found mid-range performers were more accurate than high and low performers [14].

A variety of possible factors affecting SA accuracy have been identified. Assessment task complexity and familiarity is one possible reason since students are often more

accurate in their self-assessment for less complex or more familiar tasks [15]. The Dunning-Kruger Effect mentions association with ability. People with lower ability at a task overestimate their performance, while people with high ability at a task underestimate their performance, because the less competent lack necessary meta-cognitive skills to recognize their incompetence [16]. Self-esteem also plays a role in affecting SA accuracy. High self-esteem students may overestimate themselves while low self-esteem students tend to underestimate themselves [17]. Cultural factors cannot be ignored either. In cultures valuing modesty, people might underestimate their abilities, while in culture promoting confidence, people may overestimate their abilities [18].

2.3 Focus of the Current study

The current study, therefore, tries to investigate the accuracy of self-assessment by exploring the consistency between self-assessment ratings and external ratings by teachers as well as the reasons for any discrepancies. The study first invited engineering students and their instructors to independently rate the students' spoken English proficiency using Oral Expression Scales of the "China's Standards of English Language Ability" (CSE), which is the first authoritative standard of English proficiency levels particularly designed for English language learners in China, whose counterpart in Europe is CEFR (Common European Framework of Reference for Languages). Students' self-assessment results were then compared to teachers' evaluation results and speaking test scores to investigate the correlation among these data. 23 students and 6 teachers whose results had significant differences were then invited to a semi-structured interview to explore any underlying reasons.

The objective of the study was to reveal the accuracy level of students' self-assessment and the reasons behind overestimation or underestimation in their self-assessment. The findings aim to guide engineering students towards a more comprehensive and accurate reflection of their English speaking proficiency levels, and thus help them develop appropriate autonomous plans for English speaking learning. Therefore, the specific research questions are:

1.To what extent are students' self-assessment ratings accurate in contrast to teachers' evaluation ratings and students' speaking test scores on the English course?

2.What are the possible reasons for inaccurate evaluations?

3 Model for Evaluating Self-assessment Accuracy Based on Correlation and Sentiment Analysis

3.1 Participants

The participants of this study were all the 498 first-year second-semester engineering students enrolled in a UK-China joint education program at a prestigious "Double First-Class" university in China. All the courses of the programme are delivered in English. The sample size encompasses students with a wide range of English speaking proficiency levels, laying a solid foundation for data analysis.

The study also invited 19 teachers to conduct evaluation of students' speaking proficiency levels, all of whom have over three years of experience in assessing English speaking abilities on standardized tests. Among them, 11 teachers have more than five years of such experience.

3.2 Assessment Instruments

The main instruments in this study are a student self-assessment questionnaire based on the China's Standards of English Language Ability (CSE) Self-assessment scale for Oral Expression, and a teacher evaluation questionnaire based on the CSE Overall Oral Expression Scale. The two scales assess the same constructs for oral expression abilities but the former one is designed for self-assessment and the latter one is for teacher evaluation. Both scales have 9 levels in total, but the chosen levels for this study are level 3 to level 7. This is because this range not only includes the general intermediate levels (level 4 to level 6) where the English proficiency levels of first-year non-English majors typically fall [19], but also includes level 3 for beginners and level 7 for advanced students to cover particularly exceptional and weaker student populations. The student questionnaire comprises the student's name, gender, class, descriptors from the self-assessment scale, and self-assessment levels. The descriptors for students are presented in both Chinese and English to guarantee accurate comprehension by the students. Similarly, the teacher evaluation questionnaire includes the teacher's name, years of experience in teaching and assessments, the student's name, and the teacher's rating of the student's speaking proficiency level.

The second instrument is the English course where students are taught 3 times per week in small classes of no more than 20 students and in a very interactive and communicative way. They also had a 1-1 tutorial every 3 weeks with their class teacher. All the classes and tutorials were conducted in English. Teachers, therefore, have substantial opportunities to observe a student's overall speaking performance to form an accurate understanding of his/her proficiency level.

The third instrument is the speaking assessment on the English course taken by students in the same time period. The test, including academic group discussions and academic presentations tasks, was developed and reviewed by internationally recognized experts in Academic English with over two decades of experience in teaching and research in the field. The test has a total score of 35 points with 15 for group discussions and 20 for academic presentations. All student's performances were recorded and archived for quality assurance. All the raters attended standardization training to ensure consistent understanding of the marking criteria and final scores for each student were moderated by language assessment experts before they can be released. The researchers calculated the overall reliability coefficient for the overall test as well as the reliability coefficients for the two tasks. The results revealed that the overall Cronbach's Alpha coefficient for the full oral test was 0.902, with the Cronbach's Alpha coefficients for the group discussion, academic presentation, and overall oral scores being 0.926, 0.797, and 0.812 respectively. This indicates the oral test has very high reliability and can be used for subsequent data analysis.

3.3 Model for Evaluating Accuracy of Self-assessment Based on Correlation and Sentiment Analysis

Data Collection: The study started with a thorough training on the use of the relevant CSE scales for both teachers and students, including the objectives behind the development of the CSE scales, contexts for their use, typical characteristics of CSE levels 3-7 descriptors, and differences between these levels. The next step was inviting students and their class teachers to independently evaluate students' English speaking proficiency levels based on CSE scales for oral expression. Students' self-assessment was based on their own reflection, while teachers' evaluation was based on their general observation of students' performance in class activities on the English course. Following that, students' scores for the speaking test on the English course were collected in the same time period in order to eliminate the impact of time on students' speaking proficiency levels. The last step was to select 23 students and 6 teachers whose scores had significant differences to have semi-structured interviews in order to investigate the underlying reasons. Each teacher was interviewed 3-4 times, each time about one student's speaking performance.

Therefore, data included 498 self-assessment ratings, 498 corresponding teachers' evaluation ratings, 498 speaking test scores, 23 students' interview texts and 23 corresponding teachers' interview texts.

Construction of the Model: A model of statistical analysis was constructed in order to evaluate the accuracy of self-assessment and answer the research questions. The model consists of two stages: correlation analysis using SPSS and sentiment analysis using machine learning (Fig. 1).

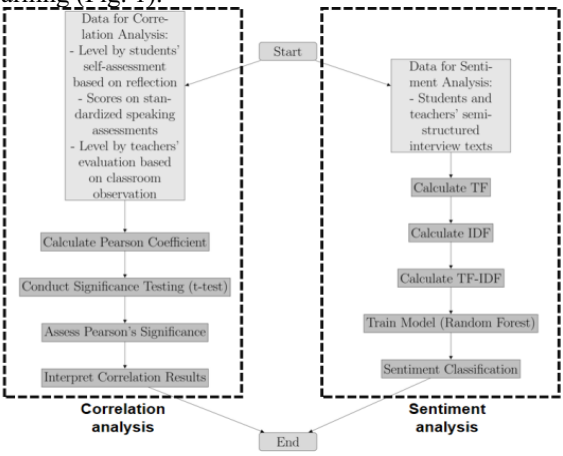

Fig. 1. Model of evaluating self-assessment accuracy

Correlation analysis: In the first stage, with the collected quantitative data, the Pearson correlation coefficient which serves as a crucial metric in quantifying the linear relationship between two variables was calculated. This coefficient varies from -1 to +1, indicating different correlation strengths and directions:

+1: A perfect positive linear relationship.

-1: A perfect negative linear relationship.

0: No linear correlation.

The process of evaluating differences between speaking test scores and self-assessment levels algorithm based on Correlation Analysis is as follows:

(1) Formula and Calculation: The Pearson correlation coefficient, represented as *r*, is computed using the following equation:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

In this formula, *n* represents the number of data points, with *x* and *y* as the variables under consideration. The symbol $\sum$ is used for summation.

(2) Significance Testing: Significance testing in statistical analysis, especially in correlation studies, is pivotal. The t-test is a prevalent method used to ascertain if the correlation between two variables significantly deviates from zero.

(3) Calculation of Significance Test: The significance of Pearson's coefficient is assessed using the formula:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

Here, *r* is the correlation coefficient, *n* is the sample size, and *t* represents the t-statistic. Degrees of freedom are computed as $n-2$. The p-value corresponding to this statistic can be determined through a t-distribution table or statistical software.

(4) Concept of Two-Tailed Significance: Two-tailed significance testing in hypothesis analysis is a method that assesses both potential directions (positive and negative) of a correlation. It primarily tests the null hypothesis of no correlation ($r = 0$).

(5)Interpretation of Results: In two-tailed tests, the p-value is doubled to reflect both tails of the distribution. A p-value lower than the significance level (commonly 0.05) suggests a statistically significant correlation, either positive or negative. Conversely, a higher p-value indicates inadequate evidence to assert a significant correlation.

Sentiment analysis: Sentiment analysis, a vital component of Natural Language Processing (NLP), leverages techniques from natural language processing, text mining, and computational linguistics. Its primary role is to discern, extract, and quantify subjective elements, particularly emotions, within textual data. The methods range from basic dictionary-based to advanced machine learning and deep learning strategies.

In this article, considering the number of samples and the length of the texts, we have chosen a machine learning approach for conducting sentiment analysis on semi-structured interview texts. The core of the machine learning method is building models based on data for prediction or classification. Sentiment analysis methods based on machine learning transform text data into features and use these features to train classifiers. The process of conducting sentiment analysis using machine learning methods, utilizing TF-IDF for feature extraction, and training the model with the Random Forest algorithm includes the following steps:

(1)Calculate Term Frequency (TF) which means a word's occurrence rate within a document.

$$\mathrm{TF}(t,d) = \frac{\text{The number of words appearing in document } dt \text{ The total number of times}}{\text{document } d \text{ Number of words}}$$

(2)Calculate Inverse Document Frequency (IDF). This involves measuring how much information a word provides, i.e., whether it is common or rare across all documents.

$$\mathrm{IDF}\ (t,D) = \log\left(\text{Total number of words included } t\right)$$

(3)Calculate TF-IDF which means the product of TF and IDF, used to reduce the impact of common words and enhance the importance of significant words.

$$\mathrm{TF-IDF}\ (t,d,D) = \mathrm{TF}\ (t,d) \times \mathrm{IDF}\ (t,D)$$

(4) Random Forest Algorithm: Random Forest is an ensemble learning algorithm that improves the overall model's accuracy and stability by combining the predictions of multiple decision trees.

Decision Tree: Features are split based on information gain. The feature for splitting that maximizes information gain is chosen.

$$IG\left(D_p, f\right) = I\left(D_p\right) - \sum_j \frac{N_j}{N_p} I\left(D_j\right)$$

Randomness: Randomness is introduced in tree construction, with each tree using a randomly selected subset of features.

Ensemble: Each tree independently makes predictions, and the final outcome is the average or majority vote of all tree predictions.

(3) Sentiment Classification: The Model is applied at this stage. Inputting features are transformed by TF-IDF into the trained Random Forest model for sentiment prediction.

(4) Prediction: The model outputs probabilities for each category, with sentiment classification based on these probability values (e.g., positive, negative, or neutral).

Throughout this process, TF-IDF is used for extracting text features, transforming raw text data into a format that can be processed by the model; the Random Forest algorithm is used for learning from these features and conducting sentiment classification. This method combines the effectiveness of TF-IDF in handling text data and the powerful performance of Random Forest in classification tasks.

Since the interview content in the current study is in Chinese, the Chinese sentiment analysis library SnowNLP is used to conduct sentiment analysis of each student and teacher's interview content. The sentiment polarity ranges from [-1, 1], where -1 represents completely negative, and 1 represents completely positive. In other words, the higher the sentiment analysis value is, the more positive the reflected content is; conversely, the lower the sentiment analysis value is, the more negative the content is.

4 Results and Discussion

4.1 Research Question 1

To what extent are students' self-assessment ratings accurate in contrast to teachers' evaluation ratings and students' speaking test scores on the English course? This question can be answered from the correlation test results obtained through bivariate correlation analysis. Table 1 and Table 2 show the correlation coefficients between self-assessment levels and levels by teachers and between self-assessment levels and speaking test scores are 0.379 and 0.345, respectively, and the significance for both is less than 0.01. That means there is a moderate correlation between self-assessment and teachers' evaluation and between self-assessment and speaking test scores. Table 3 shows the correlation coefficient between speaking test scores and levels by teachers' evaluation is 0.808, and the significance is less than 0.01, which means there is a strong correlation between the two.

However, bivariate correlation analysis alone cannot prove the causal relationship between CSE self-assessment levels and teacher evaluation levels. Therefore, partial correlation analysis was conducted to eliminate the impact of speaking scores on the two.

The partial correlation test of self-assessment levels and levels by teachers is shown in Table IV. After excluding the impact of CSE self-assessment and teacher evaluation levels on speaking test scores, we obtained the correlation between CSE self-assessment and teacher evaluation levels, and the significance is less than 0.001. There is a strong correlation between the two, but the correlation coefficient is as low as 0.180, which means that the correlation between CSE self-assessment levels and levels by teachers' evaluation is not statistically significant, and their direct correlation coefficient (0.379) is not accurate.

Table 1. Correlations between Self-assessment Levels and Levels by Teachers

		CSE Level by Ts	CSE Self-assessment Level
CSE Level by Ts	**Pearson Correlation**	1	.379**
	Sig. (2-tailed)		<.001
	N	498	498
CSE self-assessment Level	**Pearson Correlation**	.379**	1
	Sig. (2-tailed)	<.001	
	N	498	498
**. Correlation is significant at the 0.01 level (2-tailed).			

Table 2. Correlations between Self-assessment Levels and Speaking Test Scores

		CSE Self-assessment Level	Speaking Test Scores
CSE self-assessment Level	Pearson Correlation	1	.345**
	Sig. (2-tailed)		<.001
	N	498	498
Speaking Test Scores	Pearson Correlation	.345**	1
	Sig. (2-tailed)	<.001	
	N	498	498
**. Correlation is significant at the 0.01 level (2-tailed).			

Table 3. Correlations between Speaking Test Scores and CSE Levels by Teachers

		Speaking Test Scores	CSE Level by Ts
Speaking Test Scores	Pearson Correlation	1	.808**
	Sig. (2-tailed)		<.001
	N	498	498
CSE Level by Ts	Pearson Correlation	.808**	1
	Sig. (2-tailed)	<.001	
	N	498	498
**. Correlation is significant at the 0.01 level (2-tailed).			

Table 4. Partial Correlations between CSE self-assessment Levels and CSE Levels by Ts (Excluding the Impact of Speaking Test Scores)

Control Variables			CSE self-assessment Level
Speaking Test Scores	**CSE self-assessment Level**	Correlation	1.000
		Significance (2-tailed)	.
		df	0
	CSE Level by Ts	Correlation	.180
		Significance (2-tailed)	<.001
		df	495

In order to further explore the relation between students' speaking test performance and their self-assessment levels, the study classified students into 3 groups: low performers who scored at least one standard deviation below the mean, high performers who scored at least one standard deviation above the mean, and medium performers who fell between the other two [14]. It can be seen from Table V that high-performers tended to underestimate their speaking level, low-performers tended to overestimate their speaking level, and medium-performers slightly tended to underestimate their speaking level. The general trend of underestimation was also shown from T-test

results. In general, students' self-assessment levels were significantly lower that those given by teachers (M_Ts=4.82, M_Sts=4.45, t=6.2405, p-value<.05).

Therefore, in response to research question 1, though there was a moderate correlation between self-assessment and teacher assessment, self-assessment levels were not accurate when compared with teachers' evaluation levels or speaking test scores. The general trend of under-estimation was noticed, particularly for high-performing and medium-performing students.

Table 5. Count of Estimations by Performance

Performance Category	Total	No. of accurate estimations	No. of under-estimations	No. of over-estimations
High	95	25	66	4
Low	84	34	10	40
Medium	319	137	134	48

4.2 Research question 2

Discussion of Sentiment Analysis Results.

A total number of 23 students and 4 teachers were interviewed. The interview questions were set out in Table VI. It can be seen from Table VII that among the students selected for interview, 17 (74%) underestimated their levels while 6 (26%) over-estimated. Categorized from their speaking test performance, there were 4 high performers (17%), 14 medium performers (61%) and 5 low performers (22%).

Table 6. Interview Questions for Students and Teachers

Interview Questions for Students	S1: Why did you rate your speaking at a certain level? What speaking abilities do you think you have achieved? S2: Why do you think you are not the XX level given by your teacher? S3: How would you rate your overall English speaking level in the speaking test of the course? Why? S4: Are you confident in your English speaking learning? S5: How was your experience of using the self-assessment rating scale?
Interview Questions for Teachers	T1: Why did you assign a certain CSE level to a certain student's speaking? What level of ability do you think this student has achieved? T2: What do you think of the students' self-assessment level? Do you think it is accurate (over- or under-estimated)? From your observation, how do you think this student performs in speaking in classroom activities or after-class assignments? T3: Do you think the student's performance in the speaking test is consistent with your observation of his/her classroom performance or after-class assignments?

The results of sentiment analysis for students and teachers are shown in Table VII and VIII, respectively. Regarding interview Question S1, i.e. students' understanding

of self-assessment levels and own speaking abilities, the sentiment analysis results show that about 69% (score range 0.6753 to 1.0000) of students can describe their abilities relatively clearly while the other 31% (score range 0.0022 to 0.3438) thought their speaking levels were low and had a relatively negative evaluation of themselves. Compared with teachers' response to Question T1, i.e. understanding of the levels they gave to a certain student and students' speaking abilities, the results of all teachers fell within [0.6349, 1.0000], which indicates teachers were generally clearer and more positive than students about the speaking abilities each student has achieved.

Regarding Question S2 "Why do you think you are not the XX level given by the teacher?", about 9% of the students (scores between 0.1275 and 0.3257) were confused about their performance ratings and did not understand the content of the rating scale clearly, while about 91% of the students (scores between 0.7275 and 1.0000) rated themselves more positively and had a better understanding of their ability levels. Regarding Question T2 for teachers "What do you think of the students' self-assessment level? Do you think it is accurate (over- or under-estimated)? From your observation, how do you think this student performs in speaking in classroom activities or after-class assignments?", the results demonstrate that about 48% of teachers' interviews (scores between 0.0136 and 0.5911) believed that students' self-assessment was relatively negative and that students underestimated themselves to a large extent, while around 52% (scores between 0.6062 and 1.0000) of teachers' interviews believed that students' self-assessment levels were more in line with their true levels. When these interview results were combined with a comparison of self-assessment levels, levels by teachers and speaking test scores, it was shown that students who were confused about their performance ratings tended to underestimate their own ratings, and these students were generally medium performers. Teachers also believed underestimation mostly occurs among medium performers who tended to adopt a more conservative grading method for themselves due to misunderstandings of the content of the grading standards, resulting in underestimation.

Regarding Question S3 "How would you rate your overall English speaking level in the speaking test of the course? Why?", only 9% of the students (scores between 0.4228 and 0.6849) rated their English speaking level as low, while about 91% of students (scores between 0.8748 and 1.0000) evaluated their speaking level more positively and thought their overall speaking level is relatively good. Regarding Question T3 for teachers "Do you think the student's performance in the speaking test is consistent with your observation of his/her classroom performance or after-class assignments?", about 9% of the teacher interviews (scores between 0.3570 and 0.4080) thought that students' test performance is different from their classroom performance, but about 91% of teachers (scores between 0.6382 and 1.0000) believed students' test performance is basically consistent with their classroom performance observed. When these interview results were combined with a comparison of students' self-assessment levels with levels by teachers and students' speaking test scores, it was found that students who had a low evaluation of their English speaking ability, indicating lack of confidence, would underestimate their own scores, but their actual scores were at a medium level.

In response to Question S4 "Are you confident in your English speaking learning?", about 53% of the students (scores between 0.1931 and 0.5990) expressed lack of

confidence, while just under half (scores between 0.6164 and 1.0000) were confident. When considerations were also given to their self-assessment levels, it was found that all these unconfident students tended to underestimate their speaking proficiency levels which were actually between medium and high performance while only some of the confident students underestimated. This shows a certain link between underestimation and low confidence levels.

Table 7. Results of Sentiment Analysis for Students

Student	S1	S2	S3	S4	C1*	C2*	C3*
ST1	0.9763	0.9937	0.9999	0.8964	3.0000	under	Medium
ST2	0.9999	0.1275	0.9990	0.7989	1.0000	under	Medium
ST3	0.0053	0.9989	0.9986	0.5810	2.0000	under	Medium
ST4	0.9057	0.9643	0.9028	0.2932	1.0000	under	High
ST5	0.0022	0.9937	0.9999	0.6964	4.0000	under	High
ST6	0.9706	0.7275	0.9990		-2.0000	over	Low
ST7	0.3439	0.9989	0.9986	0.9810	-1.0000	over	Low
ST8	1.0000	0.9961	0.9954	0.1932	2.0000	under	High
ST9	0.0022	0.9997	0.6850	0.5930	3.0000	under	Medium
ST10	0.9999	0.9839	0.9995	0.3802	3.0000	under	Medium
ST11	0.8723	0.9953	0.4228	0.6900	2.0000	under	Medium
ST12	0.9789	0.9999	0.8767		2.0000	under	Medium
ST13	0.0084	0.9999	0.9385	0.5697	3.0000	under	Medium
ST14	0.9994	0.9818	0.9996	0.5697	2.0000	under	Medium
ST15	0.9727	0.3257	0.9989	0.5959	2.0000	under	Medium
ST16	0.9994	0.9993	0.9979	0.8796	1.0000	under	Medium
ST17	0.9988	0.9860	0.9575	0.5752	3.0000	under	Medium
ST18	0.9458	0.9458	0.9924	0.8959	-1.0000	over	Medium
ST19	0.0113	0.9995	0.9709	0.5991	4.0000	under	High
ST20	0.6753	0.7665	0.9954	0.6165	-1.0000	over	Low
ST21	0.2763	0.9936	0.9990	0.9873	-1.0000	over	Low
ST22	0.9990	0.9936	0.9811	0.5766	2.0000	under	Medium
ST23	0.9573	0.9911	0.8749	0.9761	-1.0000	over	Low

*C1: Discrepancy between self-assessment levels and levels by teachers

*C2: Over- or under-estimation

*C3: Performance Category: Medium Performer, high performer, low performer

Table 8. Results of Sentiment Analysis for Teachers

Student	T1	T2	T3	C1	C2	C3
ST1	0.9640	0.8360	0.9945	3.0000	under	Medium
ST2	0.9954	0.1008	0.9964	1.0000	under	Medium
ST3	0.9969	0.4081	0.9990	2.0000	under	Medium
ST4	0.9936	0.5159	0.8946	1.0000	under	High
ST5	1.0000	0.5257	0.9957	4.0000	under	High
ST6	0.9677	0.6062	0.7741	-2.0000	over	Low
ST7	0.9786	0.6300	0.3570	-1.0000	over	Low
ST8	1.0000	0.1102	0.9514	2.0000	under	High
ST9	0.9715	0.9977	0.9958	3.0000	under	Medium
ST10	1.0000	0.9981	0.9955	3.0000	under	Medium
ST11	0.8772	0.9263	0.9961	2.0000	under	Medium
ST12	0.9982	0.9654	0.6383	2.0000	under	Medium
ST13	0.9967	0.0136	0.4079	3.0000	under	Medium
ST14	0.9992	0.7014	0.6939	2.0000	under	Medium
ST15	0.9946	0.3297	0.7279	2.0000	under	Medium
ST16	1.0000	0.1880	1.0000	1.0000	under	Medium
ST17	0.9998	0.5751	0.9934	3.0000	under	Medium
ST18	0.6349	0.9934	0.9996	-1.0000	over	Medium
ST19	0.9980	0.5911	0.9299	4.0000	under	High
ST20	0.9730	0.6973	0.7482	-1.0000	over	Low
ST21	0.9979	0.9152	0.9947	-1.0000	over	Low
ST22	0.7552	0.1756	1.0000	2.0000	under	Medium
ST23	0.9808	0.9694	0.9936	-1.0000	over	Low

Evaluation of the Sentiment Analysis Model.

Leave-One-Out Cross-Validation (LOOCV) was used in this study to evaluate the performance of the sentiment analysis model since it is an effective way of evaluating machine learning models, especially when the amount of data is not very large. Wilks' Lambda is an important measure in LOOCV whose main purpose is to evaluate whether the differences between different groups are significant. The smaller the value is, the greater the difference between groups is. It is also used for hypothesis testing to determine whether there are significant differences between different groups. As a multivariate analysis tool, it can also consider the relationship between multiple variables, providing a more comprehensive perspective than univariate analysis.

Through LOOCV and based on the analysis of the interview texts and speaking test scores, we can see from Table IX that the overall Wilks' lambda is 0.616, which means

that the differences within groups of accurate, over- and under-evaluations are not significant, that is, there is a strong correlation between the students' evaluation of their own speaking proficiency and the content of their interviews. Specific to each question (Table X), it is found that students' optimism about their evaluation of own speaking levels is significantly positively correlated with questions S4, has a certain positive correlation with questions S2 and S3, and has a weak negative correlation with question 5. Combined with the interview content, we can know that the more accurate the students' understanding of their own abilities was and the more confident they were in their English speaking learning, the more optimistic they would be about self-assessment. The stronger the students' awareness of the difference between their self-assessment levels and the levels by teachers, and the more positive their comments were about their usual performance, the more optimistic they would be in their self-assessment. However, when students had a higher level of recognition of the CSE self-assessment rating scale, they would be more likely to underestimate their own performance.

Table 9. Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.616	8.003	5	.156

Table 10. Standardized Canonical Discriminant Function Coefficients

	Function
	1
S1	.066
S2	.291
S3	.290
S4	.981
S5	-.052

Response to Research Question 2.

What are the potential reasons for inaccurate evaluations? It can be seen from the above analysis that underestimation mainly occurred among medium-performers, and the possible reasons were students' lack of confidence in their speaking abilities and English language learning, confusion about their own speaking abilities, and lack of a clear understanding of the CSE rating scale.

5 Conclusions

The study was to investigate the accuracy of self-assessment and possible reasons for inaccuracies using correlation and sentiment analysis. Results show that there was a moderate correlation between self-assessment and teacher assessment, and a general trend of underestimation in student self-assessment when compared to teacher

evaluation or speaking test scores, particularly for high-performing and medium-performing students. This is consistent with the literature. It was also found that the possible reasons for underestimation were that students' lack of confidence in English language learning, confusion about their own speaking abilities, and lack of a clear understanding of the CSE rating scale. These findings also have valuable implications on teaching and learning. Though the accuracy of self-assessment is not validated, it could still be used as a formative assessment tool to trigger active learning and reflection on own abilities and improve assessment literacy through using rating scales or standards. It can also help teachers identify students who tend to be less confident in their learning, and thus design ways in teaching to boost their confidence. Future studies could further investigate the effectiveness of self-assessment as a tool for improving reflection abilities and assessment literacy, or explore reasons for underestimation with a larger sample size or with a population having diverse cultural or educational backgrounds.

Acknowledgment

The paper was jointly funded by the National Education Examinations Authority Ministry of Education in People's Republic of China and British Council (Project No. EARG2020008).

References

1. D. Boud, "The role of self-assessment in student grading," Assessment & Evaluation in Higher Education, vol. 14, no. 1, pp. 20-30, 1989.
2. P. R. Pintrich, "A conceptual framework for assessing motivation and self-regulated learning in college students," Educational Psychology Review, vol. 16, no. 4, pp. 385-407, 2004.
3. D. J. Nicol and D. Macfarlane-Dick, "Formative assessment and self-regulated learning: A model and seven principles of good feedback practice," Studies in Higher Education, vol. 31, no. 2, pp. 199-218, 2006.
4. H. Andrade and A. Valtcheva, "Promoting learning and achievement through self-assessment," Theory into Practice, vol. 48, no. 1, pp. 12-19, 2009.
5. D. Boud and N. Falchikov, "Aligning assessment with long-term learning," Assessment & Evaluation in Higher Education, vol. 31, no. 4, pp. 399-413, 2006.
6. T. El-Maaddawy, "Enhancing learning of engineering students through self-assessment," *2017 IEEE Global Engineering Education Conference (EDUCON),* Athens, Greece, 2017, pp. 86-91, doi: 10.1109/EDUCON.2017.7942828.
7. D. J. Nicol and D. Macfarlane-Dick, "Formative assessment and self-regulated learning: A model and seven principles of good feedback practice," Studies in Higher Education, vol. 31, no. 2, pp. 199-218, 2006.
8. P. Black and D. Wiliam, "Inside the black box: Raising standards through classroom assessment," Phi Delta Kappan, vol. 80, no. 2, pp. 139-148, 1998.
9. H. Andrade and A. Valtcheva, "Promoting learning and achievement through self-assessment," Theory Into Practice, vol. 48, no. 1, pp. 12-19, 2009.
10. S, Traci, et al. "Self-Assessment of Knowledge: A Cognitive Learning or Affective Measure?" Academy of Management Learning & Education, vol. 9, no. 2, 2010, pp. 169–91.

11. S. Lindblom-Ylänne, H. Pihlajamäki, and T. Kotkas, “Self, peer and teacher assessment of student eSAys,” Active Learning in Higher Education, vol. 7, pp. 51-62, 2006, doi: 10.1177/1469787406061148.
12. X. Fu, S. Zhong, Z. He, J. Tang, H. Chen and D. Wei, “Research on Summative Self-assessment and Formative Assessment for Big Data General Course,” in 2021 IEEE 3rd International Conference on Computer Science and Educational Informatization (CSEI*)*, Xinxiang, China, 2021, pp. 254-258, doi: 10.1109/CSEI51395.2021.9477722.
13. S. M. Zvacek, M. de Fátima Chouzal and M. T. Restivo, “Accuracy of self-assessment among graduate students in mechanical engineering,” in 2015 International Conference on Interactive Collaborative Learning (ICL), Firenze, Italy, 2015, pp. 1130-1133, doi: 10.1109/ICL.2015.7318192.
14. S. Zvacek, M. Teresa Restivo, “Accuracy of self-assessment among graduate students in mechanical engineering,” 2015 Int. Conf. on Interactive Collaborative Learning (ICL), 2015, doi: 10.1109/ICL.2015.7318192.
15. J. Andrade and G. Cizek, “Students as the definitive source of formative assessment: Academic self-assessment and the self-regulation of learning,” in Handbook of Formative Assessment, Routledge, New York, NY, 2010, pp. 90-105.
16. J. Kruger and D. Dunning, “Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments,” Journal of Personality and Social Psychology, vol. 77, no. 6, pp. 1121-1134, 1999.
17. E. Falchikov and D. Boud, “Student self-assessment in higher education: A meta-analysis,” Review of Educational Research, vol. 59, no. 4, pp. 395-430, 1989.
18. H. H. Chang, L. C. Lin, S. N. Kim, and F. D. Solovyanchik, “Cultural influences on the assessment of students’ learning in mathematics: A perspective from multiverse thinking,” Bolema, vol. 31, no. 57, pp. 1-16, 2017.
19. J. Liu, Sha. Wu, Study on China’s Standard of English Language Ability. BJ: Higher Education Press, 2019.