Journal of Pedagogical Research
Volume 5, Issue 4, 2021
https://doi.org/10.33902/JPR.2021474154

Research Article

Feasibility of using comparative judgement and student judges to assess writing performance of English language learners

Alper Şahin

Atılım University, Turkey

This study aims to identify how feasible it is to use comparative judgement (CJ) and student judges to assess the writing performance of English language learners. For this purpose, 35 paragraphs written by the students who were enrolled in a freshman Academic Writing course at a semi-private university located in the Turkish Republic of Northern Cyprus were selected and uploaded to http://www.nomoremarking.com website. Ten instructors of the Academic Writing course and 112 students taking the course volunteered to participate in the study. Then, the students were taken into 5 groups according to their writing performance level. In total, around 350 comparisons were done by each group. The results suggested that it could be feasible to use CJ to assess short writing performance like paragraphs if the instructors were experienced and trained. Moreover, instructors liked CJ and described it as a more practical, easier, fairer, faster, more enjoyable way of marking student papers. The students also liked CJ and it was also found that students who were high achievers in paragraph writing might be used to mark student papers through comparative judgement as long as they were trained.

Keywords: Comparative judgement; Computer-assisted writing assessment; Student judges; Assessing writing; Paragraph writing

Article History: Submitted 29 March 2021; Revised 7 November 2021; Published online 28 December 2021

1. Introduction

Assessing student performance effectively is one of the prerequisites of language teaching because it is desired that language learners move their language learning to performance level in order to be able to present their learning. At this stage, assessing student performance accurately is critical (1) to pinpoint student performance validly and reliably and (2) to inform the teachers and/or administrators about how their program and/or students are doing in terms of performance level English proficiency. In English language teaching, performance is mostly measured through two productive skills: speaking and writing.

Assessing speaking accurately is difficult because the level of anxiety a student experiences may confound the construct being tested (Oya et al., 2004) and cause construct irrelevant variance on

Address of Corresponding Author

Alper Şahin, PhD, Department of Basic English, School of Foreign Languages, Atılım University, Kızılcaşar Street, 06830, İncek Gölbaşı, Ankara/Turkey.

alpersahin2@yahoo.com

0000-0001-7750-4408

How to cite: Şahin, A. (2021). Feasibility of using comparative judgement and student judges to assess writing performance of English language learners. *Journal of Pedagogical Research, 5*(4), 140-154. https://doi.org/10.33902/JPR.2021474154

the student scores. Apart from this, making students speak during the exam, creating context and finding authentic situations, preparing a rubric, maintaining the intra- and inter-rater reliability are just a few problematic issues to consider while assessing speaking (Madsen, 1987).

Assessing writing is also problematic in nature similarly as it necessitates the use of subjective annotations (Verhavert et al., 2019), and the use of multiple judges for marking as the student number increases. More importantly, it requires double marking if the test results will be used for high stakes purposes. Additionally, as the variability of the responses is open, this makes the marking of essays a truly complex process (Moss, 1994). Due to this somewhat problematic nature, writing assessment is handled with more care in educational organizations and is paid utmost attention. If the reliability of marking cannot be maintained and if multiple raters assess the papers with similar qualities in totally different ways, this may lead to a decrease in the reliability of the scores attained and objections to the results may be received. As can be seen, assessing writing has some challenges difficult to solve in terms of reliability and validity (Charney, 1984).

1.1. Problems of Traditional Rubric-based Writing Assessment

There are some concerns related to validity raised out of the writing assessment (van Daal et al., 2019). First of all, Humphry and Heldsinger (2014) state that the matrix system used in analytic rubrics may even constitute a threat to validity as they may cause 'pronounced rating tendencies' which can cause Halo Effect (p.253). Moreover, what constitutes a good piece of writing is an arbitrary decision developed exponentially and uniquely by each rater throughout their service as raters or teachers. It is also backed by the research (Bloxham, 2009) that judges differ in their views of good writing. This problem can be decreased using marking schemes. However, research indicates that using marking schemes can also increase the reliability concerns (Wheadon et al., 2020) because the recent trend is to make marking schemes more specific to increase reliability which turns out to be counterproductive and this negatively affects the instruction as it comes down to narrowly following what is assessed in the marking schemes (Popham, 2005). Furthermore, although using marking schemes or rubrics helps to obtain absolute scores, the raters often make comparisons between the paper being scored and the previously scored papers (Crisp, 2013). Apart from validity and reliability concerns, using marking schemes or rubrics to assess writing requires training of the raters, monitoring them, and standardization (He et al., 2013), which can be tiring. Due to these problems, alternatives have been sought for a long time. The strongest alternative to traditional rubric-based marking is Comparative judgement (CJ).

1.2. What is Comparative Judgement?

CJ is simply based on a judge's comparing two stimuli, that is, two responses to a certain task and choosing the better one. It is not just based on the decision made by one judge though. There are multiple judges and therefore there are multiple comparisons that each stimulus is taken into account. After the repeated comparisons, the judges' decisions are tallied, and the score each stimulus gets is calculated based on these tallies and the stimuli are rank-ordered according to their standard scores. This calculation is based on the law of comparative judgement which was first introduced by Thurstone (1927). Thurstone based the law of comparative judgement on people's being better at comparing two objects with each other rather than comparing them to a preset of criteria. Thurstone's law of CJ has been reformulated in the Rasch model by Brogden (1977) and Andrich (1978). It has also been introduced to educational assessment by Pollitt (2004). A detailed review of the mathematical foundations of CJ can be found in Pollitt (2012).

CJ is well known for its high reliability (Bramley & Vitelio, 2018). One can reach up to overall score reliability of .96 (Kimbell et al., 2009) and .98 (Humphry & McGrane, 2015) via CJ. This may prove that CJ reaches reliability scores which are difficult to obtain in traditional rubric-based marking (Pollitt, 2012). CJ is also beneficial because it releases the stress of marking and makes the judges focus more on their expertise, which is thought to increase validity during the marking of, for example, essays (van Daal et al., 2019).

1.3. Challenges of CJ

Apart from the advantages it presents in terms of validity and reliability, CJ can be time-consuming and tiring for judges (McGrane et al., 2018). For this reason, adaptive comparative judgement (ACJ) has been developed. In CJ, each stimulus should be paired with all other stimuli to have a ranking score. However, in ACJ, not every stimulus has to be paired with all other stimuli. With the help of computer systems, for example, if stimulus 1 is judged to be better than stimulus 2 and if stimulus 2 is better than stimulus 3, then it is known that stimulus 1 is better than stimulus 3. Therefore, ACJ doesn't let stimulus 1 and stimulus 3 be paired and compared. This brings efficiency to traditional CJ.

Pollitt (2012) asserts that ACJ also reaches a high level of reliability that cannot be attained by any other marking methods. However, a problem stated by Pollitt (2012) was that expanding CJ even in the form of ACJ to very large-scale assessments involving thousands of students and raters would be problematic due to practicality concerns. Due to the nature of CJ (and also of ACJ), the number of paired comparisons increases as the number of stimuli (e.g. paragraphs) involved in the judging process increases. This means that more raters and more time are needed as the student number increases. It may not always be easy or cost-effective to find more raters. However, including students in the marking process may be a solution. This can be advantageous in many ways. First of all, as the number of students is plenty, if they can be used for marking, a huge hurdle in front of CJ could be overcome. Secondly, it is highly common lately to include students in assessment mechanisms in the form of peer feedback or self-reflection. Using students in marking of writing performance via CJ could be a step further to this trend without harming the reliability of the scores because it is practically impossible or rather difficult to favor one single student paper in CJ. Moreover, successful integration of students to the marking process via CJ may work in favor of performance assessment if we consider the high-reliability CJ promises in performance assessment. This may mean that if the scalability hurdle of CJ could be overcome by using the students as judges, CJ might also be implemented widely for large-scale assessment.

1.4. Challenges of CJ

The motivation behind this research was the scalability hurdle of CJ. It has been thought that this study will help to overcome this issue CJ has by addressing the following research questions:

- Is it feasible to use CJ to mark paragraphs written by English language learners?
- Is it feasible to integrate students into the marking process of paragraphs using CJ?
- What do students think about their CJ experience?

2. Method

In this part of the paper, the methodology will be presented under the research design, data collection tools, participants of the study, data collection procedure, and data analysis titles.

2.1. Research Design

In this study, both qualitative and quantitative data were collected through the Likert type and open-ended items in data collection tools. According to the Model 2 of Steckler et al. (1992), in mixed method studies, the qualitative data is collected to further explain the quantitative data. In this study, the qualitative data was also gathered to support and explain the data collected quantitatively in detail. Therefore, it can be said that this study is a mixed-method study based on Model 2 of Steckler et al. (1992).

2.2. Participants

Ten instructors from the Modern Languages Department of Middle East Technical University Northern Cyprus Campus (METU NCC) and 112 students from the METU NCC who took ENG 101 course during the spring semester of 2018-2019 academic year were used. The students and instructors were informed about the aims of the study and their voluntary contribution was asked.

Then, the students were divided into 5 groups. There were three classes (sections) that participated in the study. They were labeled as Whole Class 1, 2, and 3 (WC1, WC2, WC3). There was a total of 80 students in these classes. Moreover, students from different sections of the ENGL 101 course with high paragraph writing scores (over 8 out of 10) were also invited to the study and 32 students accepted to participate and they were divided into two groups and labeled as Skilled Raters 1 and 2 (SR1, SR2). The descriptive information about the raters can be found in Table 1.

Table 1
Descriptive information about the raters

	Male	*Female*	*Age Range*
Instructors	4	6	36-55
SR1	9	7	17-22
SR2	13	3	18-24
WC1	14	16	19-23
WC2	18	9	18-26
WC3	18	5	18-23
Total	76	46	17-55

2.3. Data Collection Tools

The data collection was done using paragraphs as stimuli because they were thought to be less time-consuming for the raters and easier to evaluate by the student raters. In addition, www.nomoremarking.com [NMM] website was used as the platform for marking. There were 10 comparisons, which was suggested by Pollitt (2012), set for each student paper; thus, a total of around 350 comparisons were reached as the minimum number of comparisons per group.

NMM was a free-to-use platform for CJ at the time this study was conducted, and it is still free for personal and research use. The paragraphs were scanned and uploaded to the system. Judge names and contact information were entered into the website. Judges received a unique link for their judgements. The system allows the judges to have a break any time and continue even after days. When the judges start their comparisons, they see two papers on the screen at the same time. It is possible to zoom a part of a paper or read the papers one by one on the screen for a larger view. The simple action that should be done by the judges was to choose the better one between the two given paragraphs presented on the screen.

A survey was prepared for both the students and the instructors who participated in the study to get their in-depth views on CJ. The survey for instructors was a survey with 3 background questions and 10 open-ended questions (see Appendix 1). Open-ended items were preferred in the instructor survey in order to collect meaningful qualitative data from a very limited number of participants (n=10). The students were administered a 7-item survey with a 5-point Likert scale. The student survey items were Likert type items because the number of the students was adequate to get meaningful results and the students' expertise on marking was low.

The background survey (which included age, gender, name surname, email address, course section, and department) was administered to student judges separately. The surveys were kept short deliberately to foster response rate as this study was solely based on voluntary efforts of both students and instructors and the task asked was rather challenging and time-consuming for a study with no incentives but the mere contribution to a research study. The survey items first were written by the researcher and they were given to a measurement and evaluation specialist to be reviewed in terms of psychometric qualities. Modifications were done based on their feedback. Then, two English instructors were asked to review the items in terms of linguistic quality and conciseness. The items were also modified based on their feedback. Then, both surveys were given to two individuals from the instructor and student groups and they were asked to read and comment on what they understand. At this stage, all items were found to be concise, and clear. Therefore, no change was made in the surveys.

2.4. Data Collection Procedure

35 paragraphs written by the students who took ENGL 101 course in the fall semester of 2018-2019 academic year during the midterm exam of the same course were selected to be used in the study. The paragraphs were around 150 words and on "The causes of sleeping disorders". The student names were removed from the papers and they were uploaded to the NMM site after being given a code number each.

Upon their consent, the student and instructor judges' names, surnames, and email information were entered to the NMM website and each rater received an email from the website with a unique link for their marking so that they could start marking. No training has been given to the instructors and the students to see their bare performance as in the nature of CJ studies (Pollit, 2012) and as CJ requires no training at all (Jones & Wheadon, 2015). However, all judges (including the student ones) were given a short speech on the aims of the study and how its findings could help the language assessment field. Their questions were answered after the speech. The prompt given to the raters during the comparisons was "Which paragraph do you think is better". The surveys were given to the students and the teachers immediately after the completion of their judging tasks. 6 instructors and 82 students responded to the surveys.

2.5. Data Analysis

The descriptive statistics like the total number of comparisons, the number of comparisons per judge, average time spent per judge, and comparison were taken from the NMM website. Similarly, Scale Separation Reliability coefficients were also taken from the NMM website. According to the NMM website (Barmby, 2019) SSR is calculated with the formula (1).

$$SSR = 1 - \left(\frac{RMSE}{SD}\right)^2 \quad (1)$$

Note. SD: Standard deviation; RMSE: Root Mean Squared Error

The scores obtained from NMM for each paper were compared to the original scores of the papers assigned by the course instructor (using an analytic rubric) during the academic term they were written. For this comparison, Pearson product-moment correlation coefficients were calculated (Jones & Alcock, 2014; Jones & Wheadon, 2015) using SPSS 22 (IBM Corp, 2013). Spearman rank-order correlation (rho) was also calculated. The original scores were obtained out of the marking of the course instructor deliberately without the intervention of a second marker in order to see what the result would be if the CJ were used instead of the course instructor.

3. Results

When the comparisons were completed, the descriptive details obtained from the NMM website were put into a table to compare the rater groups from many different dimensions. Table 2 shows these descriptive statistics.

Table 2
Descriptive statistics about the comparisons

	Number of judges	*Total number of comparisons*	*Number of comparisons per judge*	*Average time per judge (Minutes)*	*Average time per comparison (Minutes)*
Instructors	10	350	35	146.05	4.17
SR1	16	350	22	114.48	5.20
SR2	16	352	22	53.26	2.42
WC1	30	360	12	25.18	2.10
WC2	27	362	13.5	18.08	1.34
WC3	23	460	20	30.05	1.53

As illustrated in Table 2, the number of judges varies between 10 and 30, and the number of comparisons per judge varies between 12 and 35. Moreover, it can also be seen in Table 2 that the total number of comparisons is nearly the same for each judge in each judge group except WC3. It was deliberately set in this way to see if it would be more or less effective to increase the total number of comparisons when the student judges are involved in it. However, it didn't yield any positive or negative results except having a larger SSR.

An interesting piece of information in Table 2 is the average time spent for each comparison by the judges in each judge group. According to this information, the student judges mostly spent around 2 minutes for each comparison except SR1 and the instructors. They spent 4.17 and 5.20 minutes on average per comparison respectively. It is interesting to note that these two groups have spent nearly two times more time than the other judge groups.

As mentioned earlier, the original scores that each paper was assigned by the instructors of the course when the task was assigned to the students in the previous semester were compared to the scores given by the student and instructor judges through the NMM website. The correlation coefficients obtained from this analysis can be found in Table 3.

Table 3
Correlation between the original scores and scores generated by judge groups on NMM

		Instructors	*SR1*	*SR2*	*WC1*	*WC2*	*WC3*
Original Score	Pearson (r)	.65**	.43*	.39*	.19	.19	.19
	Spearman (rho)	.51**	.31	.45*	.11	.21	.22
Instructors	Pearson (r)		.44**	.47**	.37*	.38*	.30
	Spearman (rho)		.51**	.48**	.47**	.40*	.35
SR1	Pearson (r)			.44**	.48**	.24	.30
	Spearman (rho)			.29	.42*	.19	.15
SR2	Pearson (r)				.60**	.56**	.60**
	Spearman (rho)				.46*	.34	.54**
WC1	Pearson (r)					.47**	.68**
	Spearman (rho)					.33	.51**
WC2	Pearson (r)						.61**
	Spearman (rho)						.48**
Reliability (SSR)		.72	.64	.52	.53	.65	.58

Note. ** Correlations are significant at 0.01 level, * Correlations are significant at 0.05 level

As can be seen, the highest correlation (r=.65, rho=.51) with the original paper scores was obtained by the NMM scores out of instructor judges. The correlations obtained from comparative judgements by the Skilled Rater judge groups (SR1, *r*=.43, rho=.31 and SR2, *r*=.39, rho=.45) were relatively higher than the whole groups (WC1, *r*=.19, rho=.11; WC2, *r*=.19, rho=.21; WC3, *r* =.19, rho=.22). Moreover, there is a high correlation between the WC1, WC2, and WC3 scores and SR2 scores for both r and rho correlations. However, this practically means nothing as these score distributions are away from the original score distribution. It is also interesting to note that SR1 and SR2 scores both correlate with the Instructor scores (SR1, r=.44; rho=.51; SR2, r=.47 rho=.48) relatively higher than WC1, WC2, and WC3 scores (WC1-2-3, r=.19; WC1 rho=.11; WC2, rho=.21; WC3, rho=.22).

An interesting piece of information to consider in Table 3 is the SSR reliability coefficients obtained from each judge group. As can be seen in Table 3, the instructors and SR1 groups got higher reliability scores (SSR_{ins}=.72 and SSR_{SR1}=.64 respectively). SR2, WC1, and WC3 judge groups had similar reliability coefficients, .52, .53, .58 respectively. However, it is interesting to note that although SR2 scores had a higher correlation with the original score, the reliability coefficient was .52 and although WC2 scores had a relatively lower correlation with the original scores, the reliability score for WC2 was found to be .65. This may be because SSR is an internal consistency score. WC3's getting higher SSR than SR2 can be explained by their having more comparisons (460)

than the other judge groups. Additionally, all judges in WC3 may be marking equally bad and this may yield higher SSR rates. However, this may be misleading, and their scores may be arbitrary and inaccurate. Similarly, a judge group's SSR can be low, but this may be because of few judges who mark consistently different or with a huge difference than the group. As a result, the group's SSR decreases. This may be the case in the current situation as well.

3.1. Instructors' Perspective

Out of 10 instructors, six responded to the survey. The age of the instructors ranged between 36 and 55. The average year of language teaching experience of the instructors was 23 years and the average university-level teaching experience was 21. The instructor responses to 10 open-ended questions were analyzed qualitatively and the findings were presented under themes in the following sections.

3.1.1. Preference on holistic or analytic scoring

Most of the instructors who participated in the study were found to be fans of holistic scoring. The participant instructors stated that they found holistic scoring more practical, time-saving, and easier. One of the instructors also stated that "s/he could see the whole picture after years of experience". This was an expected finding given the average years of experience (23 years) of the participants. Experienced teachers may not want to deal with the details of the analytic rubrics and may want to grade holistically. More importantly, teachers prefer holistic rubrics (Bloxham et al., 2011)

Another instructor stated that although s/he was a fan of holistic scoring, s/he believed that analytic scoring contributes more to the process of standardization. This was a response from a relatively younger member of the instructor judges.

3.1.2. Views about the CJ on being an alternative

All the instructors who responded to the open-ended questions for the instructor judges indicated that they enjoyed the CJ experience that they had. Only one stated that "it was hard to concentrate, and it was odd to see the same paper again and again". Other than that, the instructors described it as an "interesting experience" and that they found it "suitable for experienced teachers" like themselves. One also stated that "it was not so hard, and it went smoothly"; however, "marking papers on the computer screen was a sort of challenge" for one. This may be an expected outcome given the average age of the instructors. All in all, it can be concluded that the instructors enjoyed their CJ experience.

Most of the instructors who responded to the open-ended questions endorsed the idea of the CJ's being a sound alternative to the traditional rubric-based scoring. One of the instructors stated that this was a nice scoring method, and it was not necessary to deal with "nitty gritty details of overly detailed grading criteria". Another instructor found it useful and stated that "comparative judgements seem to be an effective tool to support rubric-based scoring". S/he also stated that "comparative markings can be used as a reference point when reliability is affected because of delays and interruptions in traditional marking". Another one endorsed the CJ as an alternative method but also stated that a second stage was necessary to decide what to assign the best and the worst papers and called for a second stage where a further assessment criterion could be used. Last but not least, an instructor who endorsed the CJ stated that it could be used as long as the instructors' expectations were similar. There was also some sort of criticism against the CJ. One of the instructors didn't find the CJ useful because s/he thought it was not practical for her and s/he experienced concentration problems. Another instructor stated that this method was feasible only when the instructors had adequate writing marking experience. All in all, it may be concluded that although the instructors have some concerns regarding the practical uses of the CJ they see it as an alternative to the traditional rubric-based scoring.

3.1.3. Challenges, advantages and the disadvantages of the CJ

There were some challenges stated by the instructors about the difficulty to decide which paper was better. The instructors stated that more guidance was needed. One stated that although s/he didn't have to, s/he "kept thinking about the different aspects of scoring like content, organization, language, and their percentages." Another instructor stated that s/he marked them at different times, and s/he believed marking all of them at once would be better. Another challenge put forward was some papers' reappearing continuously and giving the judge the feeling of not being able to rate them accurately.

The advantages of the CJ are plenty according to the instructors. The majority of the instructors think that it is practical and faster to use the CJ. One of the instructors defines it as "a healthier approach to marking" and "more enjoyable for the teachers". Another instructor stated that the CJ "helps eliminate the problem of fairness especially if you are assessing too many papers". Another instructor endorsed this instructor by saying "it looks fairer while comparing two different levels of paragraphs in terms of weak and strong students. One also stated that "it is easier and quicker than traditional marking". Another instructor called it as "time-saving". Only one instructor stated that s/he "can't think of any". All in all, it can be concluded that the CJ was seen as a fairer, more practical, more enjoyable, easier, and quicker way of marking student papers by the instructors.

There were also some disadvantages stated by the instructors. One of the instructors stated that "you can't give feedback to students. In other words, they won't know their strengths and weaknesses". Another one criticized the idea of comparing two student performances with each other by saying "We've always been taught that we shouldn't be comparing student work. This may be wrong". Another instructor who thinks the best paper with the highest wins will get the full score stated that "the best paper may still lack some aspects and shouldn't be given full mark". Another one based his/her criticism on the cut point and standardization by saying "I don't know how the cut-off point is determined. Which papers are below and above the threshold level? How can the judges be standardized?". These concerns of the instructors indicate that although the instructors see the CJ as advantageous in terms of marking time and effort, they put forward some disadvantages based on the after marking procedures. It is clear from their statements that the teachers have concerns regarding giving feedback to the students and justifying the assigned scores to the students. As they do not know the technical calculations behind the system, their concerns about the paper with the highest wins getting full point may be ignored because there is no such rule. The paper with the highest wins does not get the full mark in CJ.

3.1.4. The time the CJ requires

The instructors think that the CJ is faster than rubric-based marking. For example, one of the instructors said: "traditional marking is more time consuming". Another instructor pointed out that "as long as you have at least one perfect paper or a perfect sample at hand already, it can be more practical". Another one mentioned that "rubrics tend to take too much time". Similarly, two instructors also state that it "depends on the task and the criteria but it's time-consuming with a rubric as the details and specific parts will slow me down" and "I would spend more time with a rubric". These responses indicated that the CJ was clearly seen as less time-consuming than traditional rubric-based marking.

3.1.5. The tasks the CJ is suitable for

The instructors have a consensus that the CJ can be used with short writing tasks. They stated that the CJ was more suitable for "paragraphs", "interviews", "presentations", "very short, focused texts", "short paragraphs". In addition, one of the instructors stated that "it is good for summative tasks". Moreover, instructors have a consensus that the CJ is not appropriate to be used for "full essays", "longer texts", "research papers", "more argumentative papers" because "there is too much to consider". One of the instructors stated that "this is not appropriate when giving feedback to students, especially in process writing". These responses indicate that the instructors see the CJ

as a suitable marking tool for short pieces of student work. This is an expected outcome as the comparison task becomes much more complicated as the length of the stimulus increases.

3.1.6. The CJ and student judges

It is obvious from their responses that the instructors did not like the idea of asking students to mark student papers via the CJ. However, although they were against it, two of them stated that they would endorse the idea "only if they were given the specifications and guidelines beforehand" and two other instructors stated that "I would be doubtful. They should be trained beforehand carefully" and "the students who mark can identify those who wrote them."

3.2. Students' Perspective

As mentioned earlier, the students were given a 7-item survey ($n = 82$) with a five-point Likert type items. The frequency and percentages of their responses can be found in Table 4.

Table 4
Frequency and percentages of the responses to student survey

		1*	2	3	4	5
1. I feel positive about the experience I had using comparative judgement while marking student papers	*f*	*4*	*5*	*22*	*34*	*17*
	%	4.9	6.1	26.8	41.5	20.7
2. I think comparative judgements can be used instead of rubric-based marking	*f*	*9*	*17*	*20*	*26*	*10*
	%	11	20.7	24.4	31.7	12.2
3. It was difficult to complete this task for me as a whole	*f*	*23*	*26*	*14*	*14*	*5*
	%	28	31.7	17.1	17.1	6.1
4. I feel like it took a long time to complete comparative judgements I did	*f*	*16*	*16*	*25*	*19*	*6*
	%	19.5	19.5	30.5	23.2	7.3
5. I think this system has the potential to be used in marking our papers in future	*f*	*8*	*17*	*19*	*28*	*10*
	%	9.8	20.7	23.2	34.1	12.2
6. It would take longer time if I scored those paragraphs using a rubric	*f*	*11*	*11*	*20*	*23*	*17*
	%	13.4	13.4	24.4	28	20.7
7. I think the students can also be used to mark papers using this system instead of the teachers.	*f*	*18*	*10*	*17*	*25*	*12*
	%	22	12.2	20.7	30.5	14.6

Note. * 1-I totally disagree, 2-I disagree, 3-Not sure, 4-I agree, 5-I totally agree

According to the responses to the first survey item, it can be said that most of the students felt positive about their CJ experience (when "I agree" and "I totally agree" responses are combined). Although around one third of the students were unsure about their experience, it is important to note that this is the item with the highest agreement rate among the other items. The responses to this item may indicate that just like the instructors, the students enjoyed their CJ experience.

By looking at the figures in Table 4, it can be said that around half of the students endorse the idea that the CJ can be used instead of rubric-based marking. It can also be said that although a large group of students agreed with the statement, one-third of the students still disagree with it (when "I disagree" and "I totally disagree" responses are combined). This may indicate that the students are unsure about the use of CJ instead of rubric-based marking. Their concern can be not being able to get feedback from their teachers and they may not be sure as they may have no idea about the distinction between the rubric-based marking and the CJ at all.

According to the responses to the third item in the survey, it can be said that the majority of the students didn't find the CJ task difficult. Only around one fourth of the students stated that it was a difficult task for them to complete the CJ task. This may indicate that the students didn't find CJ tasks that much difficult just like their instructors.

According to the student responses to the fourth item in the student survey, it can be said that most of the students felt that the time they had spent to complete CJ task was not long. However, it

should not be ignored that around one third of the students think that it took a long time for them to complete the task. It is interesting to note that this fourth item was the item with the highest unsure response. It may be because students do not have an idea about what long means in terms of marking. Still, it is important to note that the students did not find CJ task time-consuming just like their instructors.

According to the responses to the fifth item, around half of the students think that CJ system has the potential to be used in marking the papers in future. Although around one third of the students disagreed, the responses obtained for this item indicated that students endorsed the use of CJ to mark student papers in the future just like their instructors.

The responses to the sixth item in the survey indicate that around half of the students think that using CJ to mark the paragraph takes shorter than using a rubric. This result indicates that the students think that CJ takes less time than traditional rubric-based marking like the instructors.

The responses to the last item in the student survey demonstrate that around half of the students think that the students can be used to mark student papers. This may mean that the students endorse the idea of having student judges to mark papers in CJ contrary to their instructors who had some concerns about it.

4. Discussion

In this part, results regarding the research questions will be discussed under themes. Moreover, the limitations of the study and suggestions for further research will be presented.

4.1. Feasibility of putting CJ into use

The first research question was "Is it feasible to use CJ to mark paragraphs written by English language learners?" Although this study cannot give a definitive answer to this question, it can be said that the medium level correlation coefficients obtained from the original scores and the CJ scores indicated that the CJ had the potential to be used by the departments instead of the classroom assessment if the judges were trained. In addition, CJ is liked by the instructors probably due to its holistic nature. It is important to note that the instructors described CJ as a fairer, more practical, more enjoyable, easier, and quicker way of marking student papers. This is a finding that conflicts with that of Bramley et al.'s (1998) and McGrane et al.'s (2018) in which the judges found the CJ task as overwhelming and rather time-consuming. This conflict may be partially due to current participants' mostly being fans of holistic scoring and partially due to their age and experience level. Another important finding was on the suitable tasks that CJ could be put into use in English language teaching. The participating instructors stated that CJ was more suitable for short and focused performance excerpts like "paragraphs", "interviews", "presentations" and unsuitable for long pieces of writing performance like essays. When the average time spent by each instructor per judgement (4.17 mins) is considered, this is justifiable. If essays rather than paragraphs were marked through CJ, this time would be doubled or even tripled as essays are longer and more complex pieces of writing performance. This may limit the use of CJ to only paragraphs or short pieces of written or oral performance if one does not have extraordinarily patient and focused raters.

The results of the instructor survey also indicated that there might be a need to train the instructors about the CJ and to make them believe that this method works because it was obvious from the instructor survey that the instructors did not know much about how CJ worked and how the student scores were calculated.

Some concerns over the implementation of CJ were identified. First of all, it was identified that the instructors had concerns about justifying and giving feedback on the score assigned to a student paper (Jones & Wheadon, 2015) as feedback can only be attributed to analytic assessment (Sadler, 2009). These may be the primary challenges to get over before CJ can be put into use at an institution because an important part of assessing writing is to give feedback to the students after marking and justifying the score assigned to the paper. If this cannot be done, the instructors can

be under pressure and the reliability of the scores can seriously be questioned by the students. Sometimes, even a student with 97 out of 100 points objects to her score asking why she did not get 100. This may seriously increase the burden of the instructor and can damage the trust between the students and their instructors.

Another concern over the implementation of CJ is the number of instructors necessary to implement it at an institution. CJ necessitates the use of a larger number of instructors than does the traditional rubric-based assessment. Although CJ makes the writing assessment faster, easier and enjoyable, it should be noted that there were only 35 papers marked by ten instructors. If the number of paragraphs was 350, this would obviously overburden 10 instructors. It is not an improbable scenario considering that an institution with 10 instructors may easily have 200 students if each instructor teaches a single class of 20 students. This may mean that CJ has the potential to increase the burden of the instructors.

Another concern may emerge if CJ is used for the essay or longer pieces of student performance marking. Essays have multiple paragraphs and many aspects should be considered by the instructors while comparing two essays. This may slow down the comparison process and may increase the decision time between the pairs. Moreover, it may require extensive training of the instructors because, if not instructed, they probably will focus on different qualities of the essays and may utilize different strategies to decide to choose the better one. This would decrease the reliability of the judgements. Although increasing the minimum comparison from 10 to 20 as suggested by Verhavert et al. (2019) can help to increase the reliability, at least a common comparison strategy should be communicated with the raters before CJ is implemented over essays. Otherwise, it would be burdensome for the judges as they may spend more time and effort to complete the task.

4.2. The Feasibility of having Student Judges

The correlation coefficients obtained from the WC1, WC2, and WC3 indicated that using student judges as a whole without a selection criterion may not be feasible. There may be some reasons for this. First of all, it was a bit difficult to control the whole group of students while they were performing the comparisons. Some may not have taken the activity seriously. It was already expected to have lower reliability coefficients from novice assessors than expert judges (Jones & Alcock, 2014). In order to reach higher levels of reliability, more comparisons might be necessary to be done by the novice assessors (Verhavert et al., 2019). However, the correlation coefficients obtained from SR1 and SR2 student judge groups revealed some promising results. As mentioned earlier, these judge groups consisted of students who had high paragraph writing scores and thus better paragraph writing skills. As expected, they could distinguish between a good or a bad paragraph better than the other students. The correlation coefficient obtained from their judgements may seem inadequate at first glance. However, they were promising as these judges did not have previous marking experience. Moreover, from the descriptive statistics, it was seen that they already spent as much time as the instructors on each pair of comparisons. It may be thought that much better results could have been obtained if a short training had been given to SR1 and SR2 groups as it would help these students better understand why a paragraph was better than another. In this way, they also would benefit from the decreased cognitive demand that CJ necessitates for the expert raters (Liu & Li, 2012). This finding of the present study also concurs with the findings of Jones & Alcock's (2014) in which student judges performed close to expert judges as well. All in all, choosing the students who are known to be high scorers in writing and to be the ones who are aware of what a good piece of a paragraph is in CJ may be feasible and deserves to be investigated further.

The concerns of the instructors over the student judges accumulate around students' giving high scores to each other. However, they think so as they do not know how the scoring algorithms work behind the comparisons. It is not possible to favor a single paper in CJ as long as all judges are not reached and asked to favor that paper. In a class of 20 students and judges marking their

papers, this would not be possible because no one would like to favor another person's paper in such a systematic way because their papers would get a lower score in return. The only thing student judges could do would be to choose their papers as the stronger one every time they meet them while comparing papers. This would not be enough to increase the score of that paper though if at least half of the judges did not do the same for the same paper. Therefore, this is a concern with a very low probability and may be ignored.

As mentioned earlier, the last item in the student survey was about the feasibility of using students instead of the instructors to score student paragraphs through CJ. 45.1% of the students endorsed this statement. It should be noted that this is a score obtained without informing the students in detail of how CJ algorithm works and how scores are calculated. The endorsement of this idea might be higher if the students were informed more and knew how the scores were calculated. In addition, it should be noted that this endorsement could increase if the students were trained and given more opportunities to practice marking through CJ. Therefore, the responses to this item could be taken as a clear endorsement to the use of students instead of instructors in marking paragraphs by the students.

4.3. Student Perspectives over the use of CJ

It is important to first note that the students took the marking activity mostly seriously without any sort of reinforcement other than supporting scientific research. This is the desired principle in the design of similar studies (Topping, 2010). Moreover, the student survey indicated that the students mostly had positive impressions out of their CJ experience. One of the indications of this is that 43.9% of the students stated that CJ could be used instead of rubric-based marking. It was also found that the students did not find their task in CJ difficult (59.7%) and it did not take so much time of theirs (49.7%). When the average time spent for marking by the student judges is analyzed, it can be seen that the average time spent by a student ranged between 24 minutes to 114 minutes. This may mean that some students could easily complete the whole task in less than 30 minutes. However, around one-fifth (23.2%) of the students thought it was difficult for them. This can be justified as some students spent around four times more than the average on the task. Therefore, it was expected that one-third of the students had some concerns in this regard.

There are other indicators of student support towards CJ based on the responses to the student survey. 46.3% of the students stated that they see a potential in CJ to be used for marking papers in the future. However, around one-third (30.5%) had concerns either. This may be due to their not knowing how the algorithm behind CJ works and how their scores are calculated. Therefore, students should be informed about these issues in detail if the CJ is to be implemented officially.

Around half of the students (48.7%) think that rubric-based marking would take a longer time. This finding is similar to the response obtained from the teachers. This may indicate that although the students are not fully aware of how long using a rubric to mark the papers would take, they may have guessed knowing the structure the rubrics have. In addition, these students may have participated in peer feedback activities at the university previously and they may know what it takes to use a rubric while marking a text. Therefore, this may be why they were in favor of CJ.

All in all, around one-third of the students had concerns regarding the use of CJ in the marking of student papers, and there was around 50% endorsement to each statement in the survey. Therefore, it can be said that the students endorse the future use of CJ.

5. Conclusion

This research aimed to investigate the feasibility of using comparative judgement and integrating the students as judges into CJ system in order to overcome the need for more judges as the number of papers to mark increases. It was found that the teachers and the students liked the idea of CJ and stated that they would like to have CJ to be in use in the future. In addition, it was found that although the students in the present study could not get correlation coefficients as high as the experts as in the previous studies in the field (Jones & Alcock, 2014; Jones & Wheadon, 2015), if the

students with high paragraph writing scores are trained, they may be used for marking papers in CJ. It was also found that both the teachers and the students thought CJ could be used for marking paragraphs in English language teaching. However, the main concern of the instructors was that CJ might not be feasible to use for longer stimuli like essays. Rather it would be more feasible to use CJ for short stimuli like paragraphs. Moreover, the instructors also had concerns about giving feedback to the students. They thought it would be difficult to justify their scores and to give feedback to the students over their scores. In addition, one-fourth of the students who had concerns were "unsure" in many of the questions. This was an expected outcome as CJ was a totally new technique to the students and it was embraced with some concerns. All in all, it can be said that the use of CJ was endorsed by the instructors and the students who participated in the study in marking paragraphs, and the students who are high scorers and who are aware of what a good paragraph looks like may be used in CJ to mark student papers.

The instructors who participated in this study were highly experienced. This may be an advantage in CJ. Therefore, it should be noted that the findings of the study regarding the instructor perspectives could be biased due to the homogeneity of the participants' experience level in marking. In addition, the original paragraph scores used to compare the scores out of CJ were obtained from a single rater (classroom teacher) and they were accepted as true scores of the paragraphs. There was no double marking made deliberately to see what the situation would be if CJ was used instead of rubric-based marking in that course. However, this may have caused errors in the scoring of the paragraphs and the correlations between the scores may be misleading. The findings should be considered keeping this limitation in mind as well.

The findings of this research study necessitated the need for further research. Firstly, the idea to use high scorer students who are aware of what a good paragraph is in marking paragraphs via CJ should be investigated more. In the present study, students were not trained deliberately to see their bare performance before training. Therefore, the key point in such a study would be to train the student judges before marking. Secondly, a study that includes less experienced instructors can be conducted to see whether their performance and their perceptions are similar to that of experienced instructors in this study. Last but not least, further studies investigating the feasibility of using CJ in essay marking would also be desirable.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573. https://doi.org/10.1007/BF02293814

Barmby, P. (2019). The three Rs: Reliability, writing and rich tea biscuits. *No More Marking Blog*. https://blog.nomoremarking.com/the-three-rs-reliability-writing-and-rich-tea-biscuits-2de57f33e754

Bloxham, S. (2009). Marking and moderation in the UK: False assumption and wasted resources. *Assessment & Evaluation in Higher Education*, *34*(2), 209-220. https://doi.org/10.1080/02602930801955978

Bloxham, S., Boyd, P., & Orr, S. (2011). Mark my words: The role of assessment criteria in UK higher education grading practices. *Studies in Higher Education*, *36*(6), 655–670. https://doi.org/10.1080/03075071003777716

Bramley, T., Bell, J. F., & Pollitt, A. (1998). Assessing changes in standards over time using Thurstone Paired Comparisons. *Education Research and Perspectives*, *25*(2), 1-24.

Bramley, T., & Vitelio, S. (2018). The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assessment in Education: Principles, policy & Practice*, *26*(1), 43-58. https://doi.org/10.1080/0969594X.2017.1418734

Brogden, H.E. (1977). The Rasch model, the law of comparative judgment and additive conjoint measurement. *Psychometrika 42*, 631–634. https://doi.org/10.1007/BF02295985

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English, 18*(1), 65-81.

Crisp, V. (2013). Criteria, comparison, and past experiences: How do teachers make judgements when marking coursework? *Assessment in Education: Principles, Policy & Practice*, *20*(1), 127-144. https://doi.org/10.1080/0969594X.2012.741059

He, Q., Anywyll, S., Glanville, M., & Deavall, A. (2013). An investigation of the reliability of marking of the key stage 2 national curriculum English writing tests in England. *Educational Research, 55*(4), 393-410. https://doi.org/10.1080/00131881.2013.844942

Humphry, S. M., & Heldsinger, S. A. (2014). Common structural design features of rubrics may represent a threat to validity. *Educational Researcher, 43*(5), 253-263. https://doi.org/10.3102/0013189X14542154

Humphry, S. M., & McGrane, J. A. (2015). Equating a large-scale writing assessment using pairwise comparisons of performances. *The Australian Educational Researcher*, *42*(4), 443-460. https://doi.org/10.1007/s13384-014-0168-6

IBM Corp. (2013). IBM SPSS Statistics for Windows, Version 22.0. Author.

Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education, 39*(10), 1774-1787. https://doi.org/10.1080/03075079.2013.821974

Jones, I., & Wheadon, C. (2015). Peer assessment using comparative and absolute judgement. *Studies in Educational Evaluation, 47,* 93-101. https://doi.org/10.1016/j.stueduc.2015.09.004

Kimbell, R., Wheeler, T., Stables, K., Sheppard, T., Martin, F., Davies, D., Pollitt, A., & Whitehouse, G. (2009). E-scape portfolio assessment: phase 3 report. *Technology Education Research Unit.* https://www.teachertoolkit.co.uk/wp-content/uploads/2014/08/e-scape_phase3_report.pdf

Liu, P., & Li, Z. (2012). Task complexity: A review and conceptualization framework. *International Journal of Industrial Ergonomics, 42*(6), 553-568. https://doi.org/10.1016/j.ergon.2012.09.001

Madsen, H. S. (1987). *Techniques in Testing*. Oxford University Press

McGrane, J. A., Humphry, S. M., & Heldsinger, S. (2018). Applying a Thurstonian, two-stage method in the standardized assessment of writing. *Applied Measurement in Education, 31*(4), 297-311. https://doi.org/10.1080/08957347.2018.1495216

Moss, P. A. (1994). Validity in high stakes writing assessment: Problems and possibilities. *Assessing Writing, 1*(1), 109-128. https://doi.org/10.1016/1075-2935(94)90007-8

Oya, T., Manalo, E., & Greenwood, J. (2004). The influence of personality and anxiety on the oral performance of Japanese speakers of English. *Applied Cognitive Psychology*, *18*(7), 841–855. https://doi.org/10.1002/(ISSN)1099-0720

Pollitt, A. (2004). *Let's stop marking exams* [Paper presentation]. IAEA Conference, Philadelphia, PA. https://www.cambridgeassessment.org.uk/Images/109719-let-s-stop-marking-exams.pdf

Pollitt, A. (2012). Comparative judgment for assessment. *International Journal of Technology and Design Education, 22*(2), 157–170. http://dx.doi.org/10.1007/s10798-011-9189-x

Popham, W. J. (2005). The instructional consequences of criterion-referenced clarity. *Educational Measurement: Issues and Practice*, *13*(4), 15-18. https://doi.org/10.1111/j.1745-3992.1994.tb00565.x

Sadler, D. R. (2009). Transforming holistic assessment and grading into a vehicle for complex learning. In G. Joughin (Ed.), *Assessment, learning and judgement in higher education* (pp.1-19). Springer. https://doi.org/10.1007/978-1-4020-8905-3_4

Steckler, A., McLeroy, K. R., Goodman, R. M., Bird, S. T., & McCormick, L. (1992). Toward integrating qualitative and quantitative methods: An introduction. *Health Education Quarterly, 19*(1), 1–8. https://doi.org/10.1177/109019819201900101

Thurstone, L., L. (1927). A law of comparative judgement. *Psychological Review, 34*, 273-286.

Topping, K. (2010). Methodological quandaries in studying process and outcomes in peer assessment. *Learning and Instruction, 20*(4), 339-343. https://doi.org/10.1016/j.learninstruc.2009.08.003

van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., & De Maeyer, S. (2019). Validity of comparative judgement to assess academic writing: examining implications of its holistic character and building on a shared consensus. *Assessment in Education: Principles, Policy & Practice*, *26*(1), 59-74. https://doi.org/10.1080/0969594X.2016.1253542

Verhavert, S., Bouwer, R., Donche, V., & De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. *Assessment in Education: Principles, policy & practice, 26*(5), 541-562. https://doi.org/10.1080/0969594X.2019.1602027

Wheadon, C., Barmby, P., Christodoulou, D., & Henderson, B. (2020). A comparative judgement approach to the large-scale assessment of primary writing in England. *Assessment in Education: Principles, Policy & Practice, 27*(1), 46-64, https://doi.org/10.1080/0969594X.2019.1700212

Appendix 1. Survey items for instructors

1.Are you a fan of holistic scoring or analytic scoring? Why? Why not?
2.What do you feel about the scoring experience you have had while judging papers via CJ?
3.Do you think CJ can be a viable alternative to traditional rubric-based scoring? Why? Why not?
4.Were there any challenges you experienced while judging?
5.What do you think are the advantages of using CJ while marking a paper?
6.What do you think are the disadvantages of using CJ while marking a paper?
7.Do you think you would spend less or more time if you marked the papers using a rubric? Why do you think so?
8.For what kind of tasks in your context do you think this technique is more appropriate? Why?
9.For what kind of tasks in your context do you think this technique is not so appropriate? Why not?
10.Do you think students can be used to compare the papers using this system instead of the teachers?