Co-evaluation of Expositive Texts in Primary Education: Rubric vs Comments

Alba Mª Mayo Beltrán
Universidad de Extremadura

María Jesús Fernández Sánchez
Instituto Universitario de Investigación y Prospección Educativa de la Universidad de Extremadura (INPEX)

Manuel Montanero Fernández
Instituto Universitario de Investigación y Prospección Educativa de la Universidad de Extremadura (INPEX)

David Martín Parejo
Colegio Salesiano Ramón Izquierdo

Co-evaluation of Expositive Texts in Primary Education: Rubric vs Comments

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Co-evaluation of Expositive Texts in Primary Education: Rubric vs Comments [1]

Alba Mª Mayo, *Universidad de Extremadura*
María Jesús Fernández-Sánchez, *Instituto Universitario de Investigación y Prospección Educativa de la Universidad de Extremadura (INPEX)*
Manuel Montanero, *Instituto Universitario de Investigación y Prospección Educativa de la Universidad de Extremadura (INPEX)*
David Martín, *Colegio Salesiano Ramón Izquierdo*

This study compares the effects of two resources, a paper rubric (CR) or the comment bubbles from a word processor (CCB), to support peer co-evaluation of expository texts in primary education. A total of 57 students wrote a text which, after a peer co-evaluation process, was rewritten. To analyze the improvements in the texts, we used a rubric that was similar to the one in the first condition. The messages and suggestions for improvement were quantified and classified according to their range, evaluative content, and rhetorical content. Lastly, the incorporation of these suggestions in the final version of the expository text was analyzed. The results showed that the evaluative comments focused mainly on pointing out, rating, or simply correcting errors. However, hardly any justification was given for such corrections, nor were there any questions or improvement alternatives recorded for other shortcomings or non-error content. The students who co-evaluated each other with a rubric wrote more comments, addressing the different rhetorical components in a balanced way, even though these comments were written in a generic way. This might be why many of them were not incorporated in the second version of the texts, where a significant improvement could be noticed, but only in the conclusion section. In contrast, the comment bubbles recorded much more specific suggestions for correction. Although there was a slightly higher percentage of modifications in the second version of those texts, it was not enough to indicate a significant improvement in quality compared to the first version.

Keywords: Collaborative writing, Co-evaluation, Rubric, Comment bubbles, Primary Education

Introduction

From the first years of primary education, students must dedicate many years to steadily improve their writing skills, not only from a grammatical point of view, but also in terms of their semantic content, rhetorical organization, and adaptation to specific audiences and communicative contexts (Bazerman, 2013; Graham, Gillespie & McKeown, 2013).

Writing is mainly a recursive activity that requires three cyclical processes: planning what is to be communicated; writing a cohesive, coherent, and linguistically correct text; and to ensure that the result is easily understood by the intended audience and adequately reflects what was intended to be said (Hayes & Flower, 1980; Kellogg, 1996). Therefore, it can be said that both the social and communicative contexts condition the goals and processes of writing (Hayes, 1996, 2006).

Despite this fact, writing activities in schools have traditionally been approached as artificial, decontextualized, or as merely individual and linear tasks, in which there is no instruction in planning before writing, nor in revising and improving drafts. As an alternative to that approach in the classroom or in what is called the "third space" (Pahl & Kelly, 2005), scaffolded and collaborative writing activities (Graham, McKeown, Kiuhara & Harris 2012) have become a necessary alternative to address this skill (and literacy in general) in a holistic way where the social dimension also has a place. By giving an active role to the interlocutor in the revision of the text, collaborative writing tasks help to raise awareness of its communicative effects among the audience (Fernández et al., 2019). When writing, this "social way of thinking" (Mercer & Littleton, 2007) immediately confronts what the writer wants to say with what the potential reader actually understands: the communicative intention versus the meaning.

With the aim to research in more depth the effects that this type of collaborative writing could lead to, it is possible to approach the different peer-to-peer evaluative suggestions, taking into account the main support which helps the writer to carry out the review (rubrics and comment bubbles in the case of this study). Therefore, an analyze and a comparison of the messages generated could be made, as well as an examination of the individual improvements followed by peer revision. Over the last two decades, several studies have demonstrated the benefits of this type of collaborative writing practice. Not only do they lead to better texts than when students work individually (De la Paz & Graham, 2002; Harris et al., 2006; Yarrow & Topping, 2001), but they also have a positive impact on writing competency (Graham et al. 2012, 2013; Guzman & Rojas-Drummond, 2012; Fernández et al., 2019). Co-evaluation and the collaborative revision of texts foster metacognitive reflection and the self-regulation of writing (Chen, Wei, Wu & Uden, 2009;

Harris, Graham & Mason, 2006; Schunk & Zimmerman, 2007). The writers do not only benefit from the evaluation they receive; they also learn while evaluating the text of their peers (van den Bos & Tan, 2019), which ultimately enhances their own self-evaluation skills (Min, 2006; Yang, 2010).

The most widespread collaborative writing activity is the evaluation of the text written by a peer with suggestions for improvement that the writer may later include in their revision. What we call co-evaluation consists of the following steps (Montanero et al., 2014): (1) individual execution of a task or a product; (2) evaluation and feedback from a peer; (3) discussion; (4) revision and improvement of the task. It is an iterative process, since students receive a new assessment (usually from the teacher) after revising their work, specifying which improvements were introduced, and explaining the reasons for rejecting certain suggestions. The last revision task can then be carried out individually or collaboratively. The latter option has the advantage of reducing the workload of the student receiving the evaluation and provides an excellent opportunity to discuss the feedback. However, extending such discussion for too long risks disrupting the assessed student's thought process during the review of his or her task (Kollar & Fisher, 2010). Another risk of this activity is associated with the discomfort many students feel when evaluating or being evaluated by their peers (Cheng & Warren, 2005; Liu, Lin, Chiu, & Yuan, 2001). Although, the training was effective when higher education students used a rubric to assess a writing task (Şata & Karakaya, 2021). Modern digital media offer new alternatives for the development of collaborative writing. These digital resources are becoming increasingly popular because they facilitate the remote revision of texts much better than writing on paper (Goldberg, et al. 2003); in addition, they allow for the evaluation process to remain anonymous (Guardado & Shi, 2007; Wadhwa, Schulz & Mann, 2006).

The most widely known of these tools is arguably the wiki. In primary education, its use has been extensively investigated in second language learning (Li, Chu, Ki & Woo, 2012; Li & Chu, 2018; Makumoto, Chu, & Li, 2013; Woo, Chu, Ho & Li, 2011). Some of its benefits are the fostering of peer discussion, teamwork, and critical thinking.

However, when students lack the sufficient digital competence, a wiki is difficult to construct and may lead to wasted time and frustration (Li et al., 2014); this,

Mayo et al., Co-evaluation in Primary Education

along with the discomfort of editing peer work (Aydin & Yildiz, 2014), can be demotivating. In addition, it has not been shown to significantly affect their writing competency (Li et al., 2012).

Other digital collaborative writing environments, such as Google Docs, bring much more simplicity and privacy to evaluative comments (Gann, 2014). This tool allows the creation of "bubbles" in which comments are inserted and linked to a fragment of the text appearing in color, which makes it easier at the time of comparing the corrections with the original fragments. Such comments can be made simultaneously by several reviewers, either synchronously or asynchronously. It was found that the automatic logging of executed changes and the possibility of retrieving previous versions of a document were perceived very positively by users. (Kessler, Bokowki & Boggs, 2012).

In addition, this digital tool can also be used as a didactic resource to improve writing competency. Zheng, Lawrence, Warschauer & Lin (2015) found that collaborative writing practices with Google Docs exerted a strong influence on both motivation and improvement in writing proficiency, even when the evaluation was done by novice writers.

Neumann & Kopcha (2019) evaluated the writing of argumentative texts prepared by 21 primary school students, which were peer-reviewed through this application. The author of each text then wrote a new version, which was again reviewed, although this time by a teacher. The results suggest that the quality of writing improved between the first and second drafts, and the same also occurred between the second and third versions of the text. Students were not only able to detect superficial errors, but also suggested improvements related to deeper aspects such as the organization of the text; yet despite that fact, the most significant improvements occurred after the teacher's revision.

Swantarathip & Wichadee (2014) used Google Docs in a university setting, observing that students who worked with this application performed better than those who evaluated each other face to face in the classroom. However, a more recent study by Woodrich & Fan (2017) found that face-to-face co-evaluations generated better results than with this digital tool. In this context, a possible obstacle could come from the evaluated student feeling some mistrust toward the evaluator's suggestions, as there is no opportunity to discuss these suggestions in person (Blau & Caspi,

2009). Sometimes, the lack of teamwork skills also translates into overly critical, irrelevant, repetitive, or unclear evaluative comments (Farahani, Nemati, & Montazer, 2019; Ishtaiwa-Dweikat & Aburezeq, 2016). In any case, at least in first language learning in primary education, very few studies have investigated the potential of this type of digital resources and their limitations in relation to other consolidated resources in the co-evaluation of texts, such as co-evaluation rubrics.

Purpose of the Study

As we have just seen, the co-writing and co-evaluation of texts has an undeniable potential for learning to write, provided that students have the appropriate training and support. Among the classic resources, rubrics have proven to have a positive effect on the structuring of co-evaluation activities, but their real impact on the improvement of texts is conditioned by various factors, which would require further research. The limited number of studies carried out in primary education have focused primarily on narrative writing and have done little to explore the qualitative aspects of the evaluative messages and suggestions for improvement between peers.

Digital co-evaluation resources are becoming more and more widespread due to their user-friendly nature. In addition, our education system has suffered exceptional circumstances as a result of the COVID-19 pandemic, which means that these resources have often become the only viable option to support collaborative learning activities. For this reason, it is surprising that they have hardly been used in research at primary and secondary education levels. Unlike rubrics and other co-evaluation records traditionally based on paper, digital tools allow the insertion of specific evaluative corrections and comments. Furthermore, research on the adaptation of this type of software in cooperative classroom activities has become even more valuable in the current situation, given the difficulty of enabling peer-to-peer collaboration while preserving social distance.

Within this framework, the research objectives of this study were the following:

1. To analyze and compare peer-to-peer evaluative messages and suggestions to improve the writing of texts during structured co-evaluation activities at primary school level, either using a rubric, or through comment bubbles in a word processor.

2. To analyze and compare the impact of these co-evaluation resources (rubrics and comment bubbles) in the revision and individual improvement of the texts.

Method

Participants

There were 57 participants in the study, from the 5th grade (10-11 years of age) of a primary school in the city of Badajoz (Spain). Two experimental conditions were established. On the one hand, co-evaluation with rubric (CR), in which students used the help of a rubrics to evaluate the writing of a partner. In the other hand, co-evaluation with comment bubbles (CCB), in which students used the support of the comment bubbles in word to write the suggestions without any other instrument to rely on. For the distribution of the students within the two experimental conditions of the study, working pairs were randomly formed and then assigned to one of the conditions in the same way. The students in both experimental conditions showed a homogeneous level of writing in the first version of the text (without significant differences), according to the quantitative assessment of their quality obtained with the rubric and the chi-squared test used to check the distribution of the two groups. There was a very equal gender representation final distribution of the participants (Table 1).

Table 1. Distribution of the sample.

Experimental Condition	Gender	Number
Co-evaluation with Rubric (CR)	Female	15
	Male	16
Co-evaluation with Comment Bubble (CCB)	Female	12
	Male	14

Two students, who did not participate in either one of the two sessions, were excluded from the research. Also, texts written by two other students who presented specific educational needs due to their lack of proficiency in the Spanish language were not taken into account either.

Data Collection Procedure

The writing, evaluation, and subsequent rewriting of the texts were carried out in two sessions of approximately 40 minutes each.

` *Pretest.* Before starting the writing activity, the teacher gave, as a review, a brief explanation of the function and structure of expository texts (10 minutes). Immediately after the initial training, students had 20 minutes to write an expository text describing their school, using a digital device (a tablet with both Pages and Google Docs).

Intervention. Once the writing of the expository text was finished, the teacher explained to the students how to use the tool to evaluate their classmate's text (rubric or comment bubbles). For this purpose, an expository text that was not part of the study was used as a model. This exercise lasted approximately 15 minutes.

The students were randomly distributed according to the experimental conditions, and the texts were exchanged between the members of each pair. Each student individually evaluated the text written by his or her partner (10 minutes), without taking into consideration the grammar and spelling.
In the CR condition, students relied on the rubric to carry out the evaluation, noting the level of performance that the partner had achieved for each criterion, and writing some qualitative messages in the comment section.

In the CCB condition, students who used comment bubbles in the review options of Pages or Google Documents, identified the parts of the text to be improved and added some comments.
Afterwards, the texts were swapped again along with the rubrics or comment bubbles, depending on the experimental condition they belonged to. Students discussed the evaluative comments with their peers and added suggestions that made them easier to understand. This process lasted approximately 10 minutes.

Posttest. After the evaluation, the students had 20 minutes to rewrite the expository text on the same file, trying to improve it based on the comments included in the rubric or in the comment bubbles.

Data Analysis

Analysis of co-evaluation messages. The evaluative messages and suggestions for improvement recorded during the co-evaluation process were classified and

quantified according to their length, evaluative content, and rhetorical content. For this purpose, a system of categories created in a previous research on the evaluation of narrative texts (Fernández, Lucero & Montanero, 2016) was adapted to fit the specific features of the expository texts (Table 2).

Table 2. System of categories of assessment messages in expository texts.

Dimension	Category	Definition
Range	Global	Feedback on the whole text
	Local	Feedback on one or a few fragments of the text
Evaluative content	Signals	Verbal expressions in which only a fragment of the text is identified as right or wrong, without contributing anything else (for example, spelling mistakes are underlined without correcting them).
	Grade	Numerical, graphical (upward arrow, letters representing values on an ordinal scale), or verbal expressions involving only a judgment of quality. They can have a positive valence (such as numerical expressions above 70% of the scale used), intermediate valence (such as the letter R or numerical expressions between 50-70% of the scale), or negative valence.
	Question	Symbols or interrogative verbal expressions of doubt or in which clarification is requested.
	Specific correction	A graphic or verbal expression that, in addition to signaling an error, provides a specific instruction or a better alternative (includes the necessary deletion or addition of a fragment).
	Generic correction	Verbal expression that identifies a deficiency or error that is repeated (without pointing it out precisely in the text) and provides a generic instruction for improvement.
	Extension-alternative	Verbal comment that does not identify an error but suggests, exemplifies, or directly provides an alternative or non-essential extension of the text.
	Justification	Verbal commentary explaining or discussing an alternative (includes normative comments).
	Other comments	Other verbal comments, such as warnings or reinforcement measures.
Rhetorical content	Introduction	Feedback on introductory elements of the text, such as the importance of the topic, the objective (there may be some questions), and anticipation of the sections to be discussed.
	Ideas	Feedback on topics or ideas in the text. It includes the assessment of the clarity and order in which ideas are expressed, and the interest they generate.
	Support-details	Feedback on the justification or support of ideas through mechanisms such as examples.
	Conclusion	Feedback on the aspects that should appear in the conclusion of the text, such as the summary of the main ideas in relation to the initial objective.
	Formal aspects	Comments on sentence construction, punctuation, and consistency of vocabulary.
	Other	Comments on other specific meta-textual aspects, such as handwriting or margins.
	Non-specific	Feedback that is global, cross-sectional, or in which no specific rhetorical content is expressed. It does not include those annotations that are not evaluative in nature.

Adapted from (Fernández et al., 2016)

Analysis of text modifications. Finally, the incorporation of the suggestions for improvement in the final version of the expository text was analyzed, taking into account whether the suggestions were implemented, not implemented, or did not apply (because they involved comments of an emotional nature or not related to the text).

Quality assessment of the expository texts. To evaluate the quality of expository texts, a descriptive ordinal scale was created ad hoc. The rubric, very similar to the one used by the students during the co-evaluation activity, consisted of a total of 5 evaluation criteria, each of them operationalized in 4 levels of achievement (Table 3).

Each level of execution received the following scores: 0 point (level 1), 0.75 point (level 2), 1.5 point (level 3), and 2 points (level 4). In case the text did not meet the requirements of a certain level, it received the score of the next lower level. For example, one of the students named the topic of the text and posed a preliminary question (elements of level 3 in the criterion "introduction" of the rubric); however, she did not comment on other elements of level 3 in this criterion, such as the importance of the topic and the objective of the text. Therefore, her score for the criterion "introduction" was that of level 2.

To calculate the reliability of the rubric, after a brief training in its application, 36 texts produced by the students were corrected and analyzed separately by two of the researchers. The reliability index obtained, using the Kappa-Cohen method, was higher than 0.80 ($p < 0.01$) in all criteria of the rubric (Table 4).

Table 3. Rubric for evaluating expository texts.

Criterion	Level
1. Introduction	There is no introduction.
	Only the topic is stated.
	The topic, its importance, and the objective are stated (or there is a question).
	The topic, its importance, the objective, and the sections to be discussed are stated (or there is a question); or prior knowledge to understand them is stated.
2. Ideas	The information is difficult to understand or contains significant errors.
	Only one idea is explained.
	There are two or more ideas, although they are disorganized or not all well explained.
	There are two or more ideas well explained and organized.
3. Support	The main ideas are difficult to understand or contain significant errors.
	The important ideas are not explained with different words.
	The important ideas are explained with different words.
	The important ideas are explained with different words and examples are provided.
4. Conclusion	There is no conclusion.
	The conclusion consists of a single sentence or repeated idea.
	The most important ideas are summarized, but the conclusion is difficult to understand.
	The most important ideas are summarized, and the conclusion is easy to understand.
5. Formal aspects	Some sentences are difficult to understand.
	Most of the sentences can be understood, but the vocabulary is very poor.
	The sentences are easy to understand, there is a wide range of vocabulary, but there are few punctuation marks (or almost all of them are commas).
	Almost all sentences are easy to understand, well punctuated, and there is a wide range of vocabulary.
Observations and suggestions	

Table 4. The reliability index in the criteria of the rubric obtained through Cohen's Kappa coefficient.

Criterion	Introduction	Ideas	Support	Conclusion	Formal aspects
Cohen's Kappa coefficient	0.84**	0.92**	0.91**	0.94**	0.82**

Note: Statistically significant differences: (*) p <0.05; (**) p <0.01.

Results

Co-evaluation Messages

Frequency and nature of peer evaluation comments. The results of this analysis show that 92% of the subjects who reviewed texts with the support of a paper rubric (CR) suggested changes through unlinked evaluative comments at the end of the rubric, in addition to marking the corresponding levels of achievement for each criterion. In total, an average of 7.1 evaluative comments per student was recorded.

In contrast, 77.3% of the students provided suggestions through comment bubbles (CCB) by inserting comments in bubbles linked to specific fragments of the text. The remaining students did not register any comments. Overall, we found an average of 6.1 evaluative comments per student.

Range of peer-to-peer evaluative messages. Table 5 shows that the highest percentage of evaluative messages provided by students in the CCB condition referred to local aspects of the text (83%). Specific corrections suggesting the improvement of some formal aspects of the text ("*in number 1020 you are missing a period*") are quite frequent; however, most of them refer to the clarity and order of ideas.

When compared with the CCB condition, students belonging to the CR condition used a higher percentage of qualitative suggestions in reference to the text as a whole, with a predominance of signals referring to the comprehension of ideas ("*it is not easy to understand*"), semantic richness ("*the vocabulary is very poor and hard to understand*"), and grammar ("*you forgot commas and periods*").

Evaluative content. Table 5 shows that students in the CCB condition dedicated more than half of the messages to specific corrections (56.6%), providing very specific indications such as: "*the tilde (´) is missing in Ramón*". In contrast, generic corrections accounted for only 9.4% of the messages in this condition. 15.1% of the qualitative suggestions from students in the CCB condition were signals, while those in the CR condition indicated many more mistakes (31.2%). For example, one of them noted, regarding a fragment of the introduction: "*you should add more information*". Some comments such as extensions or alternatives (CCB: 3.8%; CR: 1.3%) were rarely used in both condition, and the justifications and questions were not used at all.

As for the ratings, about 13% of the messages in both conditions were verbal expressions that entailed a judgment of quality (CCB: 13.2%; CR: 11.7%). Regarding the valence of the ratings of the CCB condition, 100% were positive (such as "*I see it as well graded*"), while in the case of the rubric 75% were positive and 25% were intermediate (such as "*not bad*").

Rhetorical content. The comment bubbles of the CCB condition mainly focused on ideas and details, as well as on formal aspects and punctuation marks. An example of this is found in the messages "*you have to put more commas*" or "*instead of a period I would have put a comma*". It is worth noting the absence of comments of a semantic-lexical nature in this condition.

Although the CR condition mainly registered messages on general ideas and formal aspects, the percentage of messages is more balanced in the different categories of the table. For instance, 11.7% of the messages referred to the improvement of the conclusions section, converging in the same percentage (11.7%) the messages referred to the need to elaborate on certain ideas ("*you can add other things, such as, for example, the languages studied at school*") or vocabulary. An evaluator from the CR condition also made the following assessment to a classmate: "*the vocabulary is really poor, and it is very difficult to understand*". As in the other experimental condition, the students-evaluators had some difficulties in making comments that would lead to an improvement in the "introduction" section.

Table 5. Frequency and nature of evaluative comments.

Dimension	Category	CR	CCB
Range	Global	**34 (44.2%)**	**9 (17%)**
	Local	**43 (55.8%)**	**44 (83%)**
Evaluative content	Signals	24 (31.2%)	8 (15.1%)
	Rating	9 (11.7%)	7 (13.2%)
	Question	0 (0%)	0 (0%)
	Specific correction	**24 (31.2%)**	**30 (56.6%)**
	Generic correction	13 (16.9%)	5 (9.4%)
	Extension-alternative	**1 (1.3%)**	**2 (3.8%)**
	Justification	0 (0%)	0 (0%)
	Other comments	6 (7.8%)	1 (1.9%)
Rhetorical content	Introduction	8 (10.4%)	4 (7.5%)
	Ideas	**15 (19.5%)**	**23 (43.4%)**
	Details	9 (11.7%)	8 (15.1%)
	Conclusion	**9 (11.7 %)**	**3 (5.7%)**
	Formal aspects	19 (24.7%)	8 (15.1%)
	Other	4 (5.2%)	1 (1.9%)
	Non-specific	13 (16.9%)	6 (11.3%)
Total		220 (100%)	159 (100%)

It should also be noted that a high percentage of comments did not express a specific rhetorical content in both conditions (11.3% and 16.9%). Examples of such messages were: "*keep it up*" or "*you have to try harder*". Finally, it is necessary to highlight the very low use in both conditions of messages referring to other specific rhetorical aspects not mentioned in the previous categories, such as the margins, handwriting, or cleanliness.

Text Improvement

Implementation of evaluative messages. In order to study the improvements introduced between the first and second texts in both experimental conditions, a count was made of the changes that each participant's second text had undergone with respect to the first one, and these were then analyzed qualitatively. Figure 1 shows how the authors of the text implemented the changes suggested by the evaluators.

The results of this analysis show that almost all students who received suggestions through the rubric incorporated some suggestions for improvement (96%). In contrast, in the CCB condition, only 79.2% of students incorporated changes in at least one of the evaluation criteria.

The frequency of evaluative messages that students implemented in the final version of their texts is very similar in both experimental conditions (CCB: 52.8%; CR: 51.9%).

Regarding the rhetorical content, Table 6 shows that in the CCB condition, comments aimed at supporting the ideas in the text were the ones that prompted the most changes, while in the CR condition, the changes were rather focused on formal aspects. In this latter condition, it is worth noting the high percentage of changes prompted by comments

referring to the ideas (20%), the introduction (15%) and the conclusions (15%).

The suggestions that were not taken into account by the authors accounted for about one third of the total in both conditions (CCB: 34%; CR: 28.6%). It was particularly difficult for students to implement changes related to the order and clarity of ideas (CCB: 61.1%; CR: 27.3%).

Lastly, the comments that were classified as "*did not apply*" were those that, due to their characteristics, did not contain instructions or corrections that the author could apply to the text. Examples of these could be "*you have to try harder*" or "*everything is very good*".

Overall improvement of text quality. Table 7 shows the data related to the first and second expository texts created by the students of both experimental conditions. Their quality was qualitatively assessed by

the researchers, according to the evaluation criteria (Table 3), and without knowing which version (first or second) or experimental condition (CR or CCB) each text belonged to.

The application of the rubric shows some improvement in the scores between the first and second texts in both experimental conditions. In the case of the CR condition, better scores were observed in all evaluation criteria, although they were only significant in the specific criterion "conclusion" (t=2.29; p<0.05).

However, in the co-evaluation condition with comment bubbles (CCB) the positive changes between the two versions affected mainly the conclusion and the formal aspects. In this condition, the aspects regarding the introduction and the ideas did not show any improvement.

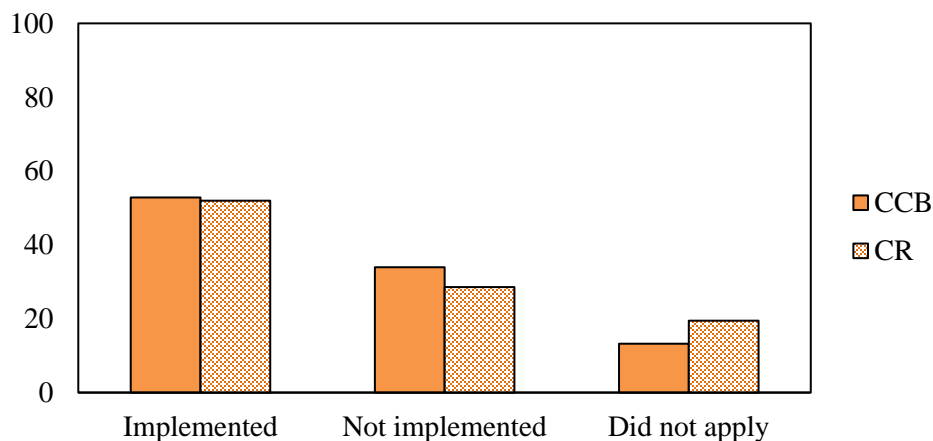**Figure 1.** Percentage of implementation of evaluative messages.

Table 7. Means and standard deviations of rubric criteria scores in the Co-evaluation with Rubric (CR) and Co-evaluation with Comment Balloon (CCB) conditions.

Text	Original text (TE1)				Final text (TE2)				Diff. TE2-TE1	
Assessment criterion	CR		CCB		CR		CCB		CR	CCB
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	Mean
Introduction	0.50	0.68	0.44	0.69	0.78	0.86	0.44	0.69	**0.28**	0.00
Ideas	1.28	0.57	1.08	0.55	1.36	0.55	1.08	0.51	0.08	0.00
Details	1.12	0.62	1.15	0.59	1.20	0.59	1.12	0.63	0.08	-0.03
Conclusion	0.21	0.57	0.19	0.39	0.62	0.79	0.30	0.53	**0.41***	**0.11**
Formal aspects	1.26	0.65	1.00	0.71	1.36	0.65	1.17	0.62	0.10	**0.17**
Total	4.36	1.90	3.87	1.69	5.32	2.17	4.11	1.89	0.96	0.24

Note: Statistically significant differences: (*) p <0.05; (**) p <0.01.

The comparison of means between the two experimental conditions shows significant differences in the wording of ideas (t=1.96; p<0.05) and total scores (t=2.22; p<0.05), in favor of the CR condition.

Discussion and Conclusion

This study compared the effects of two resources to support peer co-evaluation of expository texts in primary education: a paper-based rubric and a digital comment tool of a word processor.

Regarding the first aim of the research, the analysis of the evaluative comments reveals a certain lack of reliability in both conditions of co-evaluation, due to the small number of comments and to the nature of the evaluative content. The comments were mainly focused on pointing out, rating, or simply correcting errors. Hardly any justification for such corrections was given, nor were there any suggestions of alternatives to improve the content or organization of the text.

Clear differences were found in the type of evaluative messages generated by peer co-evaluation activities, depending on which support tool was used. In general, the use of a rubric prompted 15% more global comments on the quality of the text, mainly focused on rating it, pointing out errors, or making very generic suggestions, which were distributed over the different parts of the rhetorical structure of the text. This could be explained by the very systematic nature of the rubric, which guides students through an extensive evaluation of the most relevant criteria. Yet, the subjects clearly had difficulties in making specific suggestions for improvement beyond what was already stated in the levels of achievement of the rubric, which resulted in mostly generic messages. An important fact illustrating this point is the 30% of unsuccessful modifications generated by the rubric, which suggests that the students were unable to interpret a large part of the evaluative comments made by their peers. Several studies have shown the difficulties encountered by students when interpreting tables and other external representation systems in school tasks (Gabucio et al., 2010). The results of our study also show that, among primary school students, there might be a need to complement the use of rubrics with other resources in order to ease the transition from evaluation to help.

In contrast, the use of the "comment" tool of a word processor generated fewer and more specific evaluative messages. Although the students received the same explanation about the rubric criteria during the training activity, it seems that such knowledge was not sufficient to generate a systematic evaluation of the different rhetorical and meta-textual aspects, without the actual presence of the rubric as a guide during the co-evaluation. As expected, the "bubbles" did not express general comments, but rather pointed out and corrected specific errors. They mainly focused on semantic corrections (missing or poorly expressed ideas and information-details). This last conclusion does not coincide with previous studies, which found that peer review with digital tools such as Google Docs generated comments focused on lexical and grammatical aspects, to the detriment of organizational and semantic content (Ge, 2011; Wang, 2009). Ultimately, although the digital comment tool facilitates the insertion of concrete suggestions for improvement, it does not in itself generate a sufficiently comprehensive co-evaluation of the semantic and organizational content of the texts.

On the other hand, we know that feedback with digital media tends to be more abundant, i.e., to provide more information than the feedback on paper (Goldberg et al., 2003; Zheng et al., 2015). However, in this study the use of a rubric proved to be a resource that fostered a more abundant and balanced feedback in terms of semantic and rhetorical content, when compared with digital environments. These findings are consistent with the conclusions of MacArthur (2009), which stated that digital tools such as "comment bubbles" are not sufficient to prompt text improvements, but must be combined with instruction to improve the quantity and quality of revisions.

Regarding the second objective, the vast majority of students who were evaluated with a rubric made changes following the suggestions of their peers (96%), compared to 77.3% in the case of students who used comment bubbles. This result contradicts the findings of Chou (1999), as well as Tsui and Ng (2000), who found that less than 50% of students incorporated the suggestions received in co-evaluation situations without this type of resource. The difference could be explained precisely by the influence of comment bubbles, and especially that of rubrics, as triggers and structurers of co-evaluation messages. We know that the revision process is further enhanced when learners

have the opportunity to compare the initial version of the text with peer suggestions before deciding whether to accept or reject them (van den Bos & Tan, 2019). This seemed to be particularly true in the CCB condition.

In both conditions, some positive changes were found in matters related to ideas, although it is worth noting the high percentage of modifications regarding formal aspects in the CR condition. These results differ from what was found in a previous study in which students barely incorporated orthographic or grammatical changes (Montanero, Lucero & Fernández, 2014).

However, it cannot be concluded that the co-evaluation activity had a significant impact on the improvement of the text. Although the texts generally increased in length and the average overall assessment of their quality was higher in the second version, the difference was only statistically significant in the conclusion section of the CR condition. In addition, approximately one third of the improvements were not directly related to the suggestions received, but were triggered spontaneously during the individual phase of review.

Future studies should explore whether an increase in the number of co-evaluation activities, as well as prior training in the use of co-evaluation resources, could generate significant improvements in the texts. Training has already been identified in previous studies as one of the crucial factors of co-evaluation with rubric (Jonsson & Svingby, 2007). It then seems reasonable that such training should focus not only on the interpretation of the criteria, as was done in this study, but also on providing more concrete suggestions for improvement.

On the other hand, we also intend to study whether a combination of both co-evaluation resources, rubrics and comment bubbles, could contribute to compensate for the limitations detected and create a positive impact on peer co-evaluation, both in the improvement of the texts and in the writing skills of primary school students.

References

Aydin, Z., & Yildiz, S. (2014). Using wikis to promote collaborative EFL writing. *Language Learning & Technology*, 18(1), 160-180.

Bazerman, C. (2013). Understanding the lifelong journey of writing development. *Infancia y Aprendizaje*, *36*(4), 421-441. DOI: 10.1174/021037013808200320

Blau, I., & Caspi, A. (2009, October). *Sharing and collaborating with Google Docs: The influence of psychological ownership, responsibility, and students' attitudes on outcome quality*. Paper presented at the World Conference on E-learning in Corporate, Government, Healthcare, & Higher Education. Vancouver, Canada.

Chen, N.S., Wei, C.W., Wu, K.T., & Uden, L. (2009). Effects of high level prompts and peer assessment on online learners' reflection levels. *Computers & Education, 52*, 283–291. DOI: 10.1016/j.compedu.2008.08.007

Cheng, W., & Warren, M. (2005). Peer assessment of language proficiency. *Language Testing, 22*(1), 93–121. DOI: 10.1191/0265532205lt298oa

Chou, M. C. (1999). How peer negotiations shape revisions. J. Katchen, & Y.N. Leung (Eds.), *The Proceedings of the Seventh International Symposium on English Teaching* (s. 349–359). Taipei: The Crane Publishing Co.

De la Paz, S., & Graham, S. (2002). Explicity teaching strategies. Skills and knowledge: writing instruction in middle School classrooms. *Journal of Educational Psychology*, *94*(4) 687-698. DOI: 10.1037/0022-0663.94.4.687

Farahani, A. A. K., Nemati, M., & Montazer, M. N. (2019). Assessing peer review pattern and the effect of face-to-face and mobile-mediated modes on students' academic writing development. *Language Testint in Asia*, *9*(18). DOI: 10.1186/s40468-019-0094-7

Fernández, M. J., Lucero, M., & Montanero, M. (2016). Rojo sobre negro. ¿Cómo evalúan los maestros las redacciones de sus estudiantes? *Revista de Educación*, 372, 63-82. DOI: 10.4438/1988-592X-RE-2015-372-315

Fernández, M. J., Montanero, M., & Lucero, M. (2019). La evaluación de la competencia narrativa en la educación básica. *Revista de Educación*, 383, 85-112. DOI: 10.4438/1988-592X-RE-2019-383-402

Gann, D. A. (2014). Pedagogical Affordances Of Two Online Document Types. *Language Teacher*, *38*(4), 31.

Ge, Z. (2011). Exploring e-learners' perceptions of net-based peer-reviewed English writing. *Computer-Supported Collaborative Learning, 6,* 75–91. DOI: 10.1007/s11412-010-9103-7

Goldberg, A., Russell, M., & Cook, A. (2003). The effects of computers on student writing: A meta-analysis of studies from 1992–2002. *Journal of Technology, Learning, and Assessment*, *2*, 1–51.

Graham, S., Gillespie, A., & McKeown, D. (2013). Writing: importance, development and instruction. *Reading and Writing, 26*(1), 1-15. DOI: 10.1007/s11145-012-9395-2

Graham, S., McKeown, D., Kiuhara, S., & Harris, K.R. (2012). Meta-analysis of writing instruction for students in elementary grades. *Journal of Educational Psychology*, *104*(4), 879-896.

Guardado, M., & Shi, L. (2007). ESL students' experiences of online peer feedback. *Computers and Composition*, *24*(4), 443–461. DOI: 10.1016/j.compcom.2007.03.002

Guzman, K., & Rojas-Drummond, S. (2012). Escritura colaborativa en alumnos de primaria. Un modo social de aprender juntos. *Revista Mexicana de Investigación Educativa, 52*(17), 217-245.

Harris, K. R., Graham, S., & Mason, L. H. (2006). Improving the writing, knowledge and motivation of struggling young writers: effects of self-regulated strategy development with and without peer support. *American Educational Research Journal*, *43*(2), 295–340.

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. C.M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (s. 1–27). Hillsdale: Lawrence Erlbaum Associates.

Hayes, J. R. (2006). New Directions in Writing Theory. C. A. MacArthur, S. Graham & J. Fitzgerald (Eds.), *Handbook of writing research* (s. 28–40). New York: The Guilford Press.

Hayes, J. R., & Flower, L. (1980). Identifying the organization of writing processes. L.W. Gregg & E.R. Steinberg (Eds.), *Cognitive Processes in Writing. An Interdisciplinary Approach* (s. 3-30). Hillsdale: Lawrence Erlbaum.

Ishtaiwa-Dweikat, F. F., & Aburezeq, I. M. (2016). The Use of Google Apps to Support Teaching and Learning: A UAE Case Study. *International Journal of Web-Based Learning and Teaching Technologies*, *11*(4), 1-21. DOI: 10.4018/IJWLTT.2016100101

Jonsson A., & Svingby, G. (2007). The use of scoring rubrics: reliability, validity and educational consequences. *Educational Research Review*, *2*, 130–144. DOI: 10.1016/j.edurev.2007.05.002

Kellogg, R.T. (1996). A model of working memory in writing. C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (s. 57–71). Hillsdale: Lawrence Erlbaum Associates.

Kessler, G., Bokowski, D., & Boggs, J. (2012). Collaborative writing among second language learners in academic web-based projects. *Language Learning & Technology*, *16*(1), 91 – 109.

Kollar, I., & Fischer, F. (2010). Peer assessment as collaborative learning: A cognitive perspective. *Learning and Instruction*, *20*, 344-348. DOI: 10.1016/j.learninstruc.2009.08.005

Li, X., Chu, K., Ki, W. W., & Woo, M. M. (2012). Using a wiki-based collaborative process writing pedagogy to facilitate collaborative writing among Chinese primary school students. *Australasian Journal of Educational Technology*, *28*(1), 159-181.

Liu, E., Lin, S., Chiu, C., & Yuan, S. (2001). Web-based peer review: the learner as both adapter and reviewer. *IEEE Transactions on Education*, *44*(3), 246-251.

MacArthur, C. A. (2009). Reflections on Research on Writing and Technology for Struggling Writers. *Learning Disabilities Research & Practice*, *24*(2), 93-103.

Makumoto, M., Chu, S. K. W., & Li, X. (2013). Peer-feedback and revision process in a wiki mediated collaborative writing. *Educational Technology Research*

and Development, 61(2), 279–309. DOI: 10.1007/s11423-012-9285-y

Mercer, N., & Littleton, K. (2007). *Dialogue and the development of children's thinking: A socio-cultural approach.* London: Routledge.

Min, H. T. (2006). The effects of trained peer review on EFL students' revision types and writing quality. *Journal of Second Language Writing, 15*, 118-141. DOI: 10.1016/j.jslw.2006.01.003

Montanero, M., Lucero, M., & Fernández, M. J. (2014). Iterative co-evaluation with a rubric of narrative texts in Primary Education / Coevaluación iterativa con rúbrica de textos narrativos en la Educación Primaria. *Infancia y Aprendizaje: Journal for the Study of Education and Development, 37*(1), 184-220. DOI: 10.1080/02103702.2014. 881653

Neumann, K. L., & Kopcha, T. J. (2019). Using Google Docs for Peer-then-Teacher Review on Middle School Students' Writing. *Computers and Composition, 54*, 102524. DOI: 10.1016/j.compcom.2019.102524

Pahl, K., & Kelly, S. (2005). Family literacy as a third space between home and school: Some case studies of practice. *Literacy, 39*(2), 91-96. DOI: 10.1111/j.1741-4350.2005.00406.x

Şata, M., & Karakaya, I. (2021). Investigating the Effect of Rater Training on Differential Rater Function in Assessing Academic Writing Skills of Higher Education Students. *Journal of Measurement and Evaluation in Education and Psychology, 12*(2), 163-181. DOI: 10.21031/epod.842094

Schunk, D., & Zimmerman, B. (2007). Influencing children's self-efficacy and self-regulation of reading and writing through modeling. *Reading & Writing Quarterly, 23*(1), 7-25. DOI: 10.1080/10573560600837578

Tsui, A. B. M., & Ng, M. (2000). Do secondary L2 writers benefit from peer comments? *Journal of Second Language Writing, 9*(2), 147-170.

Van des Bos, A. H., & Tan, E. (2019). Effects of anonymity on online peer review in second-language writing. *Computers & Education, 142*, 103638. DOI: 10.1016/j.compedu.2019.103638

Wadhwa, G., Schulz, H., & Mann, B. L. (2006). Effects of Anonymity and Accountability During Online Peer Assessment. In Bruce L. Mann (Ed.), *Selected Styles in Web-Based Educational Research* (s. 300-331). Hershey: Information Science Publishing.

Wang, P–L. (2009). The application of integrated literature instruction in freshman English teaching. *Sino-US English Teaching, 6*(9), 1–11.

Woo, M., Chu, S., Ho, A., & Li, X. (2011). Using a wiki to scaffold primary-school students' collaborative writing. *Educational Technology & Society, 14*(1), 43-54.

Woodrich, M., & Fan, Y. (2017). Google Docs as a tool for collaborative writing in the middle school classroom. *Journal of Information Technology Education: Research, 16*, 391-410.

Yang, Y-F. (2010). Students' reflection on online self-correction and peer review to improve writing. *Computers & Education, 55*, 1202–1210. DOI: 10.1016/j.compedu.2010.05.017

Yarrow, F., & Topping, K. (2001). Collaborative writing: the effect of metacognitive prompting and structured peer interaction. *British Journal of Educational Psychology, 71*, 261–28. DOI: 10.1348/000709901158514

Zheng, B., Lawrence, J., Warschauer, M., & Lin, C.H. (2015). Middle School Students' Writing and Feedback in a Cloud-Based Classroom Environment. *Technology, Knowledge and Learning, 20*(2), 201-229. DOI: 10.1007/s10758-014-9239-z

Mayo et al., Co-evaluation in Primary Education

Citation:

Corresponding Author:

María Jesús Fernández Sánchez
Instituto Universitario de Investigación y Prospección Educativa de la Universidad de Extremadura (INPEX)
Spain

Email: mafernandezs [at] unex.es