British Journal of Educational Technology *Vol 51 No 5 2020* 1615–1630
doi:10.1111/bjet.12987

Controlled evaluation of a multimodal system to improve oral presentation skills in a real learning setting

Xavier Ochoa and Federico Dominguez

Xavier Ochoa is an Assistant Professor of Learning Analytics in Steinhardt School at New York University (NYU). Xavier has been an active researcher in this field of Learning Analytics, currently being an Editor-in-Chief of the Journal of Learning Analytics. His main research interest is the automatic measurement and feedback of 21st-century skills through multimodal analysis. Federico Dominguez is a professor at the Faculty of Electrical and Computer Engineering of ESPOL University and head of the Smart Environments Lab since 2014. His main research interests are IoT technologies applications in learning and working environments. Federico is a Senior IEEE member and enjoys jogging and biking in his free time. Address for correspondence: Xavier Ochoa, ALT Department, Steinhardt School, 82 Washington Square East, 7th floor, New York, NY 10003, USA. Email: xavier.ochoa@nyu.edu

Abstract

Developing oral presentation skills requires both practice and expert feedback. Several systems have been developed during the last 20 years to provide ample practice opportunities and automated feedback for novice presenters. However, a comprehensive literature review discovered that none of those systems have been adequately evaluated in real learning settings. This work is the first randomised controlled evaluation of the impact that one of these systems has in developing oral presentation skills during a real semester-long learning activity with 180 students. The main findings are that (1) the development of different dimensions of the oral presentations are not affected equally by the automated feedback and (2) there is a small but statistically significant effect of the use of the tool when a subsequent presentation is evaluated by a human expert.

Introduction

Being able to "speak in public to inform, self-express, to relate and to persuade" (De Grez, 2009, p. 5) is widely regarded as one of the fundamental competencies required by professionals in an information-based society (Chan, 2011; Riemer, 2007). Given its importance, curricular standards bodies have included oral presentation skills in their core set of requirements, eg, Common Core (Kyllonen, 2012) in the USA and the Joint Quality Initiative (2014) in Europe.

Despite the importance of being able to orally communicate with an audience, most people experience high levels of anxiety when confronted with the prospect of speaking in public (Hamilton, 2013). Fortunately, psychological studies have demonstrated that oral presentation skills are not fixed characteristics, but can be taught and learned (Fawcett & Miller, 1975) and that, in most cases, anxiety recedes with practice and experience (Miller *et al.*, 2017). Following the importance and trainability of oral presentation skills, higher education institutions have introduced several strategies to develop these skills in their students. In an extensive survey of tested existing strategies, van Ginkel, Gulikers, Biemans, and Mulder (2015) identified seven principles needed to effectively teach oral presentation skills: (1) clearly identify that the learning objective is becoming a better presenter, (2) the presentation should be aligned with the content that the student is learning, (3) provide students with observable models of peers or experts, (4) opportunities for

Practitioner Notes

What is already known about this topic

- Using multimodal sensors and AI-based algorithms for computer vision and voice processing, it is possible to build systems to provide automatic feedback for oral presentations.
- These systems accurately measure several dimensions of the quality of the oral presentation, are usually well received by presenters, and laboratory-based evaluations indicate that there are learning gains that appear after using the systems.

What does this paper adds

- Different oral presentation dimensions are affected differently by the use of the system. There are large measurable gains in looking at the audience during the presentation, while there is a negligible improvement in the avoidance of filled pauses.
- Evidence found in this paper suggests that automated feedback has a positive effect on oral presentation quality, but that the strength of this effect is small.

Implications for practice

- The current generation of automated feedback systems is better for courses where there is no allocation of resources for human-generated feedback.
- In courses where there is human expert feedback, the report produced by these systems could be used as an effective reflection-inducing artefact shared by students and experts during feedback sessions.

practice, (5) provide feedback that is explicit, contextual and timely, (6) involve peers in formative assessment and (7) facilitate self-assessment. There is evidence that these principles lead to better results in different learning contexts (van Ginkel *et al.*, 2015).

Even if the path to the development of oral presentation skills is known, its teaching and assessment are a time-consuming activity (Chan, 2011, De Grez, 2009). Principles 4 and 5, practice and feedback, usually require the involvement of additional individuals apart from the learner and take place during classroom time. As such, finding the opportunities for extensive practice and teacher-provided feedback competes with achieving other learning goals, especially when the development of oral presentation skills is integrated into regular courses as principle 2 recommends. When confronted with this dilemma, not only for oral presentation skills, but also for the teaching and assessment of most of "*21st-Century*" skills, Griffith and Care (2014) conclude that current feedback and assessment practices, while correct, are not scalable or feasible to conduct in regular classrooms and that "*new forms of data collection needed to be devised, and methods of analysing those new forms of data needed to be identified and tested*" (p. 13). In line with this conclusion, Multimodal Learning Analytics (MmLA) techniques have been proposed as a potential solution to this dilemma. MmLA, true to its origins in traditional Learning Analytics, focuses on the use of multimodal data to better understand and improve learning processes (Ochoa & Worsley, 2016). In the specific context of developing oral presentation skills, and as will be shown in the following section, MmLA has proposed solutions to both fulfil the desirable pedagogical principles (providing ample opportunity for practice and feedback) and facilitate its integration in regular courses and classrooms (reducing the amount of class time and human-resources needed).

While there are considerable examples of automated oral presentation feedback (Batrinca, Stratou, Shapiro, Morency, & Scherer, 2013; Damian *et al.*, 2015; Ochoa *et al.*, 2018; Schneider,

Börner, van Rosmalen, & Specht, 2015; Trinh, Asadi, Edge, & Bickmore, 2017), there is no strong scientific evidence that they actually help with the development of oral presentation skills or if those skills developed with the system are transferred and evident in real-world presentations (see next section for evidence). The main contribution of this work is to conduct the first large-scale randomised controlled experiment in an authentic setting to test the effect that one of these systems, RAP (Ochoa *et al.*, 2018), has on the oral presentation performance of entry-level higher education students.

Previous research

While it is clear that it is possible to build automatic systems to provide feedback to oral presentations, it is not clear if those systems fulfil their purpose of helping novice presenters to gain oral presentation skills. This section will review the existing literature to establish if and how those systems have been evaluated. These papers were found by previous experience of the authors and queries for related terms in academic search engines. While to the author's knowledge, these papers represent the most important research in the field, it is not a systematic literature review. To facilitate the summarisation of this review, previous work will be classified according to the objective of the evaluation. This review is also presented in the Table S1.

System accuracy

The most common type of evaluation performed in early systems was to test if the output provided by the system to the presenter was similar to human expert-generated feedback for the same presentation. Automatically generated estimations of different presenter's features (looking at the audience, voice volume, etc.) were correlated or compared with similar estimations made by human experts. Both Presentation Sensei (Kurihara, Goto, Ogata, Matsusaka, & Igarashi, 2007) and Cicero (Batrinca *et al.*, 2013; Wörtwein *et al.*, 2015) were only evaluated in this way. Newer systems, Gan, Wong, Mandal, Chandrasekhar, and Kankanhalli (2015), Autommaner (Tanveer, Zhao, Chen, Tiet, & Hoque, 2016) and RAP (Ochoa *et al.*, 2018), have their accuracy measured to establish if the automatically estimated values correlated with human annotations of the same features. All these evaluations have been laboratory-based because the participants were not students or the setting was a non-authentic learning activity. In general, these evaluations indicate that the automatically extracted features can be used as a proxy for an expert evaluation when the rubric is specific enough (eg, rate the eye contact with audience), but not when the evaluation made by the human expert is holistic (eg, grade the general quality of the presentation).

Perceptions about the system

Most reviewed studies (12 out of 16) have evaluated the perception that users have about the system. In these, presenters are surveyed or interviewed after using the system about several aspects related to their experience. While the type and specificity of the questions vary greatly between studies, there are some common enquiries: "Is the system useful?", "Would you use the system again?" and "Did you learn with the system?". The answers are generally positive. It is important to note that some studies (Schneider *et al.*, 2015; Tanveer *et al.*, 2015; Trinh *et al.*, 2017) use the survey to compare different types of feedback interfaces between the tools. These surveys have only been conducted in laboratory settings. Only two studies (Damian *et al.*, 2015; Schneider, Börner, Van Rosmalen, & Specht, 2017) were conducted in a real-world setting, but only with 3 and 12 participants, respectively.

Learning within the system

If the presenter uses the system on more than one occasion, or if the system provides online feedback, the measurements made by the system could be used to estimate if the presenter is

learning. Learning in this context is defined by obtaining a better score in one or more presentation dimensions or by counting the duration and frequency of mistakes. Only Logue (Damien *et al.*, 2015) and two versions of the Presentation Trainer (Schneider *et al.*, 2015; Schneider *et al.*, 2017; Schneider, Romano, & Drachsler, 2019) were evaluated in this way. In the case of Logue, the study was conducted as a lab-based controlled experiment within-subjects, where the performance was measured first without the system being active, and then, measured with the system working as designed. This evaluation found that only speech rate improves in a statistically significant way (evaluating 15 presenters). In the case of Presentation Trainer, Schneider *et al.* (2015) also used a lab-based controlled experiment with a control and an intervention group (20 participants each). This study found a statistically significant difference in the time that the presenters stay at erroneous states. Schneider *et al.* (2017), in an observation of 12 students using the system in-the-wild found that they stay less time in mistake states after a training session although no statistical test for significance was performed. In the virtual-reality version of the Presentation Trainer, Schneider *et al.* (2019) in another laboratory-based observation reported again statistically significant improvements in the same metric after three training sessions.

Learning outside the system

The most challenging evaluation for these systems is to identify if they have a measurable impact on the acquisition of oral presentation skills after using the system. Only three studies have tried to establish the actual effectiveness with different levels of success. Tanveer, Lin, & Hoque (2015) were the first to evaluate their system, Rhema. Their study consisted of a laboratory-based controlled experiment where 30 recruited participants were assigned to both control and intervention conditions in random order. The recorded videos of the presentations were then rated using a basic rubric by 10 judges. These judges were layman without expertise in oral presentations. Maybe due to these limitations, this study did not find any statistically significant difference between presentations with or without the system. In the only in-the-wild study of this category, Schneider, Börner, Van Rosmalen, and Specht (2016) found that Presentation Trainer has a statistically significant positive impact on peer-reported scores of presenters after using the system. However, this study was not controlled for external variables and included only nine students. The most recent study of this type by Trihn *et al.* (2017) evaluated the learning transfer after using the system and was laboratory-based with 12 participants and 12 judges, all of them with various degrees of experience in presentations. The study found that only a few presentation aspects were statistically significantly improved.

It is clear from the review that no study could be used as strong evidence about how useful these systems are. Most of the evaluations have centred on the accuracy of the system or the subjective perception of users. The studies that have centred on the acquisition of oral presentation skills have been mostly conducted with a small population in non-authentic, non-controlled settings. Moreover, studies that had as objective measuring the transfer of those skills to real-world situations present methodological flaws (eg, using laymen or peers as judges or using a practice session as the final measure), limiting the ecological validity of their conclusions. It is the main contribution of the present work to conduct the first in-the-wild, large-population study of the effect of an automated feedback system on the development of oral presentation skills, both within the system during practice sessions and outside the system in real presentation evaluated by human experts.

Description of the system

This section presents a brief description of the Automatic Presentation Feedback System (shortened to RAP due to its Spanish acronym) to help the reader better understand the evaluation. An

extended description of the system, technical details of its inner workings and an evaluation of the attitudes of students towards the system can be found in Ochoa *et al.* (2018).

The main component of the RAP system is the presentation capture room where users perform an oral presentation in front of a virtual audience. The hardware in the room records the presentation through three modalities: audio, video and a digital presentation file (slides). The room layout can be seen in Figure 1. The presentation slides and the virtual audience are shown in two screens opposite to one another. Embedded in the room, two small form-factor computers capture the presentation using a low-cost camera and a microphone. The camera is camouflaged within the virtual audience screen and the microphone is placed outside the field-of-view of the presenter. The presenter is recorded for 5 minutes in each practice session.

The RAP system extracts presentation features from the three modalities to generate the feedback report (Figure 2). Using the Computer Vision library OpenPose (Cao, Simon, Wei, & Sheikh, 2017), the video is analysed to extract the skeletal joints of the presenter for each video frame.

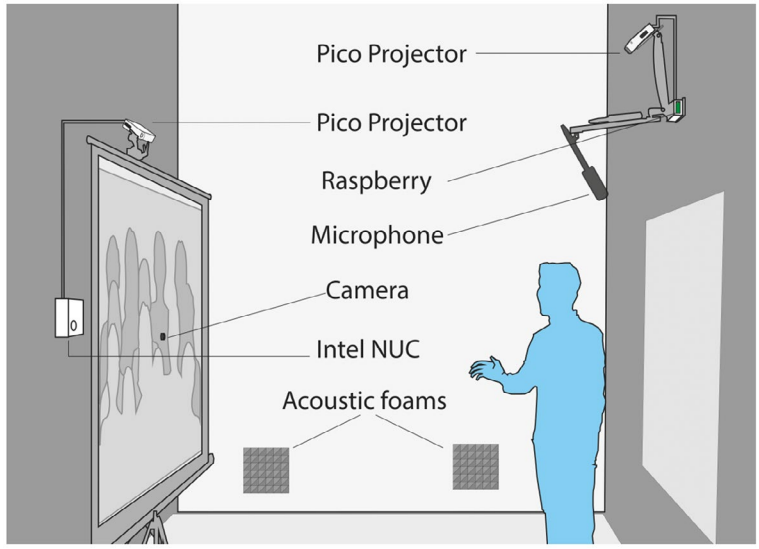

Figure 1: The RAP system's presentation room
[Colour figure can be viewed at wileyonlinelibrary.com]

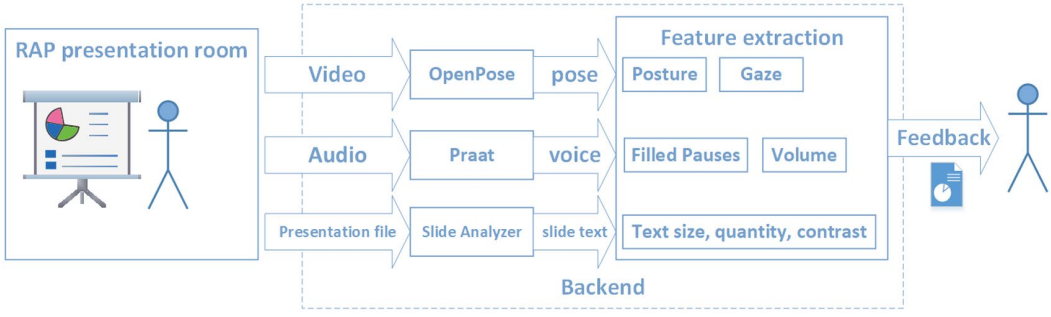

Figure 2: Recording and feature extraction on the RAP system
[Colour figure can be viewed at wileyonlinelibrary.com]

This information is used to estimate the presenter's pose and gaze direction. These estimations are then classified as correct or incorrect, depending on rules and models built with the advice of expert presentation trainers. The PRAAT speech analysis library (De Jong & Wempe, 2009) is used to extract two features from the audio recordings: voice volume and filled pauses. The value of these features is again classified as correct or incorrect depending on previously built models. Finally, a simple algorithm is used to analyse the digital file containing the slides to obtain three features per slide: font size, text length and colour contrast. The three digital file features are combined into a "slide quality" feature per slide.

After a presentation recording, the system calculates a five-level score for each one of the five features (posture, gaze, volume, filled pauses and slide quality) depending on the percentage of correct instances during the whole presentation. Using this information, the system then composes a feedback report that can be viewed by the presenter shortly after the presentation (Figure 3). The feedback contains a small evaluation of the presentation with a complete recording, an overall score and individual scores for each one of the dimensions. Additionally, for each dimension, the report includes examples (pictures or audio clips) of the presenter's correct and incorrect moments during the presentation.

The RAP system has been deployed "in-the-wild" at ESPOL where two presentation capture rooms are fully functional. This system has been in place since May 2018 and it has been under regular use by thousands of students to this date.

Evaluation methodology

As demonstrated in the Previous Research section, the impact of these systems in facilitating the acquisition of oral presentation skills has not been evaluated. To address this lack of evidence, this work seeks to answer two research questions:

RQ1: Is there demonstrable acquisition of oral presentation skills due to the use of the system?

RQ2: Do the skills gained with the system have any influence in the oral presentation performance in real scenarios as evaluated by an expert?

To be able to answer these questions with scientific rigor in a real usage scenario, a randomised controlled trial was designed and conducted with actual engineering students in the context of a communication-focused course during the second half of 2018 in the institution where RAP is in operation. From all the courses that were scheduled to use the RAP system during the second semester of 2018–2019, the course Communications II was selected. This was an opportunistic selection based on four reasons: (1) The course provided a real learning context where students' oral presentations are evaluated by an expert both in form and content; (2) focusing on students currently taking Communications II reduced the variability that would have been introduced by working with students at different levels in the Communications line; (3) some professors that teach this course had previous experience using the system; and (4) the course provided students with human formative feedback on oral presentations. The RAP system was designed to augment human feedback, not to replace it. Three professors from this course, covering a total of six sections, agreed to participate in the experiment. This agreement was needed because the experiment required extra work and some small modifications of their pre-established schedule. None of the selected professors were involved with the design of the system, but all of them have used the system before.

Students registered in these courses without previous knowledge that the experiment will be carried out. Given that the sections were scheduled for different times and that there were taught by different professors, it was expected that the population of students involved in the experiment

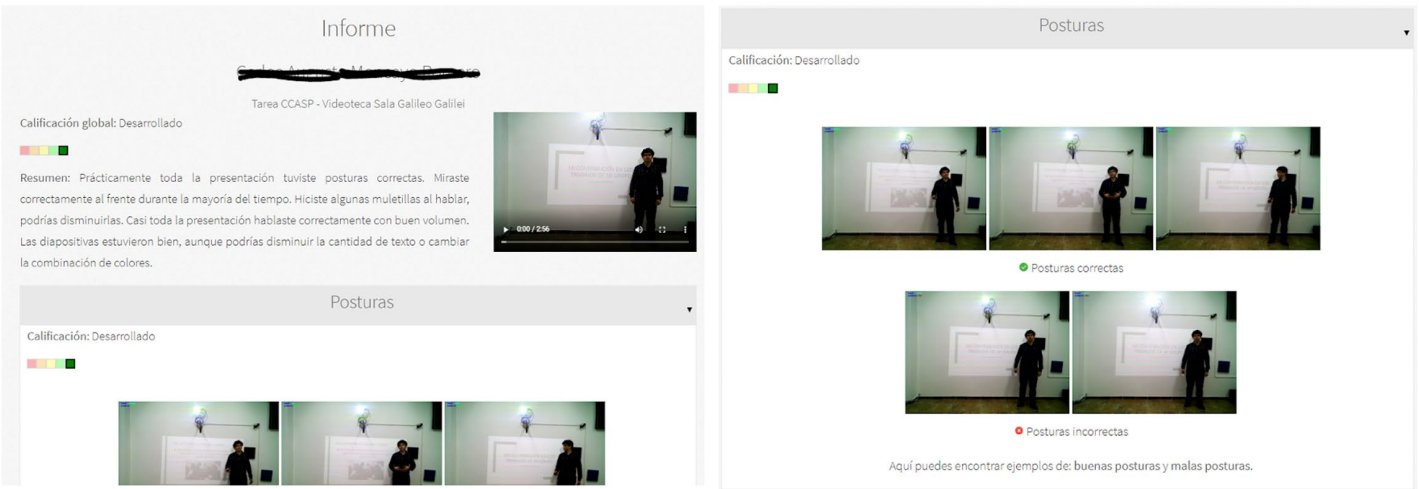

Figure 3: Example of actual automated feedback provided by the RAP system where students can review their entire presentation together with examples of their mistakes
[Colour figure can be viewed at wileyonlinelibrary.com]

are representative of the whole population of students at the institution. In total 226 students were finally registered in the six selected sections. To allow for a controlled trial, all the students were randomly assigned into two groups: 115 in a case group and 111 in a control group.

All students were given three presentation assignments during the semester (16 weeks long). The first assignment was conducted during week 2. There were 3 weeks between the first and second assignments and 10 weeks between the second and third assignments. The first and second assignments happened before the topic of oral communication was discussed in the classroom. The third assignment was part of the usual course design and took place near the end of the course. Figure 4 presents the timeline of the experimental activities. Each assignment consisted of preparing a 5-minute presentation of a current general interest topic supported by presentation slides. For the case group, the presentations of the first-two assignments were recorded with the RAP system in a predefined recording room. In the control group, the presentations of the first-two assignments were recorded by the students using their mobile phones. As students that did not receive feedback could not be required to attend the RAP rooms, one difference between the groups was the recording setting. This difference, however, should not have a major effect in the study, as literature (van Ginkel *et al.*, 2015) does not mention rehearsal context as a major influence in presentation skills acquisition. For the third assignment, all students (control and case) presented in front of their professor and classmates at the end of the semester. With this design, both groups had the same amount of practice, but only the case group received automated feedback.

Logistically, a scheduling web application was used to assign a date and time for case group students to use one of the RAP rooms. Control group students used the same application to upload a video of their recordings before the deadline. After each measurement, students in the case group were able to review their recorded presentation and the corresponding automated feedback. This feedback consisted of scores and report about four dimensions of presentation skills: maintaining an open posture, looking at the audience, voice volume and use of filled pauses. The analysis of the digital presentation files feature was not used, due to the nature of the third presentation assignment (no slides were used). Students in the control group were also able to review their uploaded video but no automated feedback was provided. Professors were asked to review the recorded videos of the second measurement and provided general feedback and comments to students via the system for both control and case groups. Table 1 presents a summary of the three measurements. The objective of this experimental design was to isolate the effect of the RAP feedback on the oral presentation skills of the students controlling for expert feedback received and practice opportunities, the two other major controllable factors that could affect the development of oral presentation skills (van Ginkel *et al.*, 2015). Initial individual ability was controlled through the random assignment of students to control and case groups.

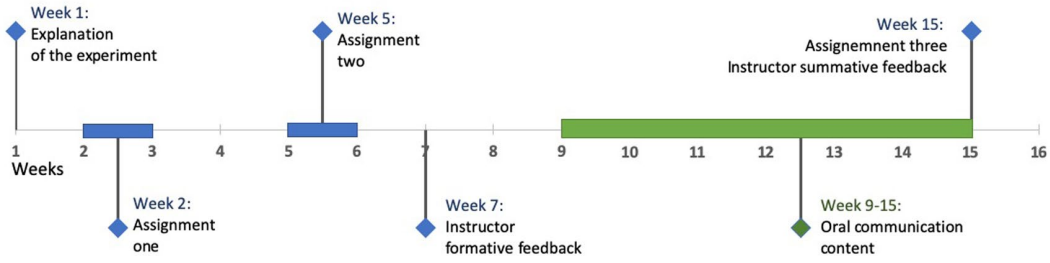

Figure 4: Timeline of the experiment during the 16 weeks of the course
[Colour figure can be viewed at wileyonlinelibrary.com]

Table 1: Distribution of presentation location and origin of feedback for case and control groups

	Presentation		*Feedback*	
	Case	*Control*	*Case*	*Control*
Measurement 1	RAP room	Mobile phone	Automatic	None
Measurement 2	RAP room	Mobile phone	Automatic and expert	Expert
Measurement 3	Classroom	Classroom	Expert	Expert

To assure that professors remained unaware of the case/group distribution of their students during the entire semester, they only reviewed the second measurement from sections where they did not teach. Additionally, they were asked not to enquire about their students' use of the RAP system during the semester to avoid accidentally learning the student's group classification.

The score for the third and final measurement, which is used to establish the effectiveness of the system, was provided by the professor in charge of each section (single-score). To assure a minimum level of concordance between the three professors, a workshop was organised to agree on a common rubric for the third measurement. This rubric was based on the indicators that Communications II professors use to rate the oral presentation of their students. This rubric can be seen in Table S2. The final score of this rubric was used as the expert evaluation of the oral presentation performance of the students. Moreover, to further ensure inter-rater agreement, the professors rated six pre-selected RAP room recordings from previous semesters using the agreed-upon rubric. The initial Cronbach's alpha was acceptable (0.82). Professors then discussed their rates until consensus was reached.

Parallel to the experimental measurements, two surveys were conducted to obtain additional information. Students in both the case and control groups were asked to fill out a survey before the first measurement. It contained the following questions (in Spanish): (1) Have you used the RAP system before? (Yes/No); and (2) How would you rate your presenter skills? (Scale 0–10). After the second measurement, students in the case group were presented with a second survey that contained the following questions: (1) What do you remember best about the RAP reports received? (Open question); (2) Based on the feedback received on previous uses of the RAP system, what do you do best? (Multiple choice + Open question); (3) According to the feedback received in previous uses of the RAP system, what should you improve on? (Multiple choice + Open question); and (4) How would you rate the feedback received so far by the RAP system? (Scale 0–5).

At the end of the semester, the data were collected from the database and stored files of the online system. These data were analysed to check the validity of the experiment and to answer the two research questions. This analysis and its results are presented in the next section.

Results and discussion

The analysis of the experimental data is divided into several subanalyses that are described in the following subsections. Following each analysis, there is a discussion of the implications of its results.

Validity of control and case grouping

The nature of the authentic setting for the experiment created some issues that needed to be considered to assure the validity of the analysis. First, some students (14) had used RAP system in other courses. Four of those students were in the case group and 10 were in the control

group. These students were not considered in the analysis. Only students that completed the three measurements were considered in the study. This rule excluded 22 students. For the analysis, 85 students remained in the case group and 95 in the control group. The total attrition was 20% of the population (26% for the case group and 14% for the control group). A visual representation of these numbers can be seen in Figure 5. The higher level of attrition in the case group can be explained by the extra work required to schedule and attend the recording session.

The self-evaluation of the abilities was used to assure that the random distribution of students into case and control created two comparable groups before and after attrition. To facilitate the analysis, the 10-point scale used in the answer was reduced to three levels: "High" [10–7], "Medium" [7,4] and "Low" [4,0]. The general distribution of responses for both groups can be seen in Table 2. Most of the students declare that they were at "Medium" level (initial and final: 64%), with a smaller number declaring "High" (initial: 24%, final: 25%) and "Low" (initial: 12%, final: 11%) levels. This distribution is similar in the initial and final sampling of both the case and control groups. A Chi-Square test (initial: $p = 0.62$, final: $p = 0.76$) confirmed that no ability level was overrepresented in any of the groups before or after attrition. Finally, A Kruskal–Wallis test was applied to determine if students with different self-declared ability levels in the case group received different scores in the dimensions measured by the RAP system. This test did not detect significant differences most probably due to the low statistical power consequence of the low number of students in the High (21) and Low (11) levels.

Learning as measured by the system

Given that the case group used the system twice, the effect of the first round of practice and automated feedback could be measured by comparing the scores obtained during the first and

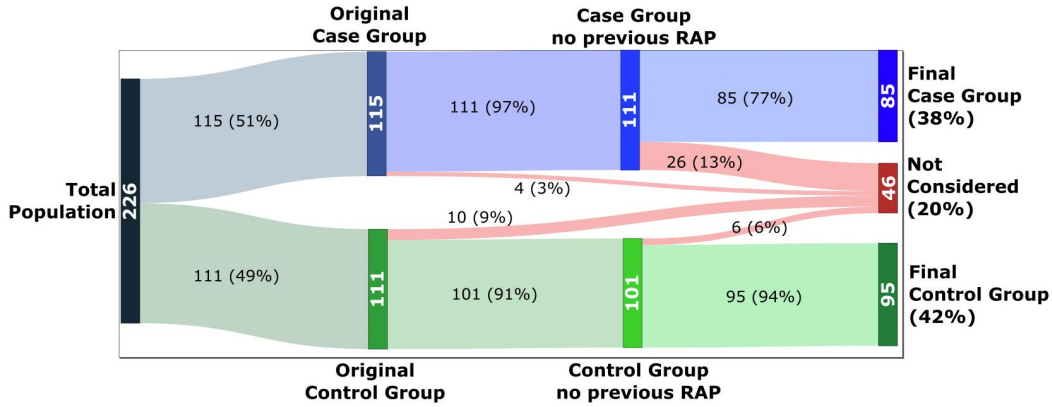

Figure 5: Division of the population between case and control group with attrition percentages causes (1) previous use of the RAP system and (2) not completing all the assignments
[Colour figure can be viewed at wileyonlinelibrary.com]

Table 2: Distribution of self-perceived ability among case and control groups from initial to final sampling

Ability\Group	*Case (Initial/Final)*	*Control (Initial/Final)*	*Total (Initial/Final)*
Low	16 (14%)/11 (13%)	11 (10%)/9 (10%)	27 (12%)/20 (11%)
Medium	73 (63%)/53 (62%)	72 (65%)/62 (65%)	145 (64%)/115 (64%)
High	26 (23%)/21 (25%)	28 (25%)/24 (25%)	54 (24%)/45 (25%)

second measurements. Figure 6a presents the distribution of scores in the first measurement for the four dimensions measured by RAP. Two of the dimensions (open posture and voice volume) present a heavily skewed distribution of values, where 94% are concentrated between 4 and 5 scores. This implies that the RAP system is not able to find students that regularly maintain a bad posture (looking to the slides, crossed arms, etc.) or that speak too softly in the population of students in the case group. The scores for the other two dimensions (looking at the audience and use of filled pauses) are more uniformly distributed, meaning that the RAP system is better able to detect errors in these areas. The same analysis was performed for the second measurement (Figure 6b). The "open posture" and "voice volume" remain heavily skewed to high scores. It can also be observed that there is an obvious increase in the "looking at the audience" scores. Additionally, there is very little visual difference in the "filled pauses" dimension. A paired sign test was conducted to accept or reject the null hypothesis that the second score is not higher than the first score. This test was selected because the data are not normally distributed and it is highly skewed. To measure the effectiveness of the use of the RAP system in students with different levels of oral presentation skills, the test was also conducted with subgroups of high performers (scores 4 or higher in the first measurement) and low performers (scores 3 or lower in the first measurement). The results can be seen in Table 3. The test found no statistically significant difference in the "open posture" and "voice volume" dimensions. The statistical analysis, however, confirmed the visual evidence that there is a significant improvement in the "looking at the audience" dimension for all the students. The effect is stronger for low performing students, where more than

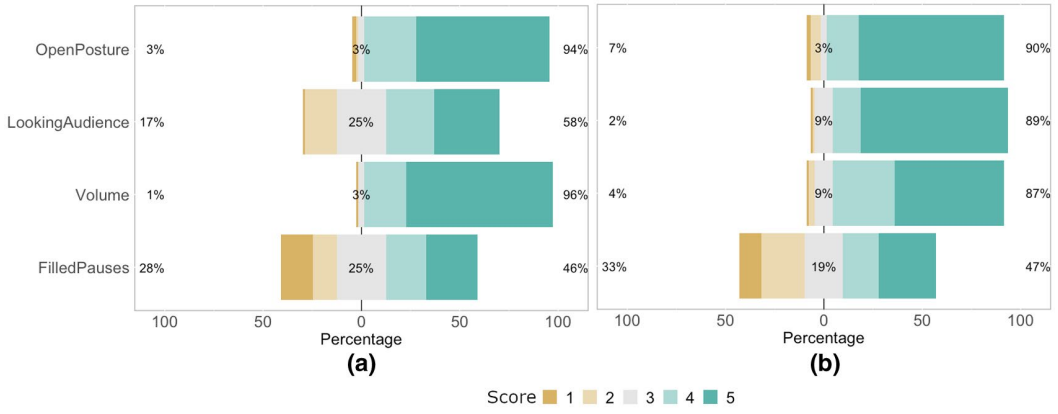

Figure 6: Distribution of the scores obtained by the students using the RAP system during the first measurement (a) and during the second measurement (b)
[Colour figure can be viewed at wileyonlinelibrary.com]

Table 3: Result of the paired sign test between the first and second RAP measurements

Dimension	*Total*	*Low Performers*	*High Performers*
Open posture	No increase	N/A (less than 10 students)	No increase
Looking at audience	***+1 median increase*** ***(p = 0)***	***+2 median increase*** ***(p = 0)***	***0 median increase*** ***(p < 0.01)***
Voice volume	No increase	N/A (less than 10 students)	No increase
Use of filled pauses	No increase	***0 median increase*** ***(p < 0.05)***	No increase

half received two additional points in the second measurement. For high performers, there is also a statistical increase but is weaker as most of them do not change their score. The only other statistically significant increase was detected for low performers in the "use of filled pauses" dimension, but with a negligible effect.

To determine if students with different levels of self-declared ability had different learning gains, the same paired signed test was conducted by ability group (High, Medium and Low). This test detected a significant difference in line with the general result only for students in the Medium group. However, no conclusion could be extracted from this result as the statistical power of the test is compromised by the low number of students in the High (21) and Low (11) groups.

These results provide an answer to the first research question (R1). This answer, however, is nuanced. There is clear and strong evidence that at least one dimension (looking at the audience) was positively affected by the use of the RAP system and its automated feedback. The feedback provided also produced a marginal reduction in the frequency of filled pauses used during the presentation for those that use them regularly (low performers). The evidence is less clear for the "open posture" and "voice volume" dimensions as the initial measurement saturated the scoring scale of the system, complicating the measurement of improvement.

Remembering the feedback

Before the third measurement, and without previous notice, the students in the case group were surveyed about which presentation dimensions they perform better and in which they need to improve the most. It was possible to select one or more dimensions, answer "I do not remember" or leave the questions blank. All the students considered in the case group answered the question about their strengths (85 students), while 89% (76 students) answered the question about their areas of improvement. Only 7% of the students declared that they did not remember the feedback.

To verify that the students indeed remembered the feedback given by the system, the percentage of students that both obtained the highest score (five) in a given dimension and answered that they were good at that dimension was calculated for both the first and second measurement scores. Results can be seen in the top half of Table 4. On average, 81% of students that received the top score in a dimension in the first measurement, report that they remember that the system told them that they were good at that dimension. This percentage increased to 84% when the second measurement is considered (Figure 4). When the score of individual students was considered, it seems that there is a high level of agreement between obtaining a high score in a

Table 4: Comparison of survey answers about remembering the feedback and actual scores obtained

Answer to Question 2	*I have a good posture (45 students) (%)*	*I look at the audience (14 students) (%)*	*I have a good voice volume (51 students) (%)*	*I do not use many filled pauses (11 students) (%)*
Score M1 = 5	94	78	86	64
Score M2 = 5	82	86	94	72
Answer to Question 3	*I need to improve my posture (14 students) (%)*	*I need to look more at the audience (26 students) (%)*	*I need to improve my voice volume (7 students) (%)*	*I use many filled pauses (28 students) (%)*
Score M1 < 5	71	73	57	96
Score M2 < 5	71	31	85	86

dimension (both in the first or second measurement) and remembering good performance in that dimension (above 80% on average). A similar calculation was performed for the dimension where students believe they need to improve based on the feedback (bottom half of Table 4). Here, an average of 74% of students that received some kind of feedback (score below five) in a dimension during the first measurement reported that they need to improve in that dimension. However, in this case, this percentage decrease to 68% for students that received feedback during the second measurement. This decrease is mainly due to the corresponding decreases for the "looking at the audience" and "filled pauses" dimensions. It seems that for dimensions in which the students improved ("looking at the audience" and "filled pauses") the students answered according to the scores received during the first feedback session. It seems that this first report was especially memorable when dealing with erroneous behaviour. The main conclusion of this analysis is that students liked and remembered the feedback they received, even from the first feedback session.

Learning as measured by an expert

The evaluation of the usefulness of the system consisted in comparing the performance of students that use the system and received its feedback report (case group) with the performance of those students that had the same opportunities for practice and the same professor-based feedback (control group), judged by a human expert during an oral presentation in a real-world learning context. The distribution of the scores for both groups is presented in Figure 7. A visual inspection of the distributions suggests that scores of students in the case group have a higher median (18) than in the control group (17). As the distribution approaches normality, usual statistics of mean and standard deviation are applicable. The control group obtained an average of 16.62 (SD = 2.28) and the case group a mean of 17.15 (SD = 1.87). Visually there seems to be a high overlap between the distributions of scores of both the case and control group. A *t*-test, however, indicates that the mean of the grades of the case group was statistically significantly higher than the mean of the control group ($p < 0.05$). The calculation of Cohen's d between

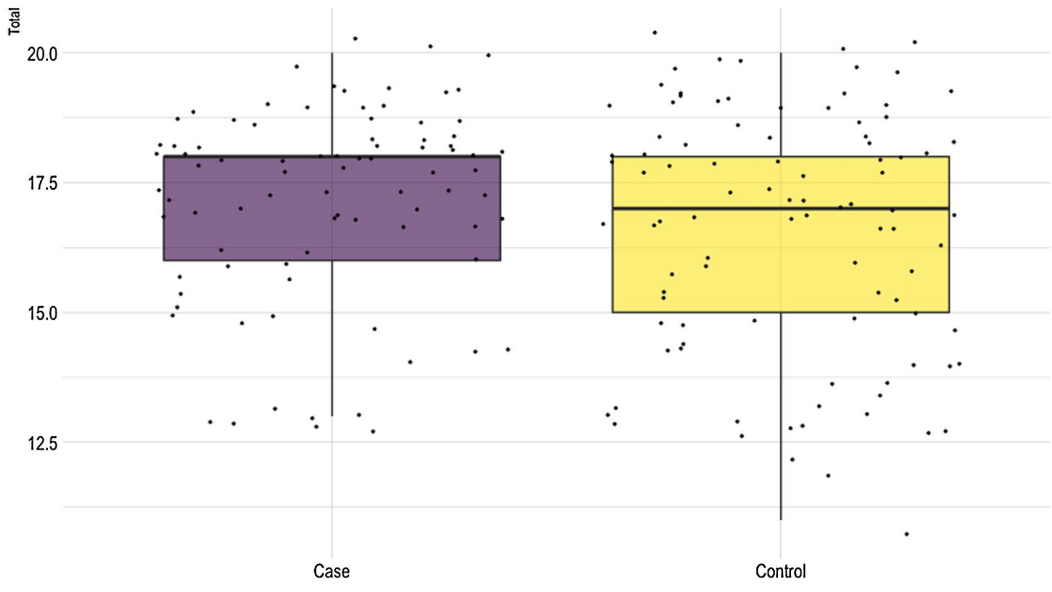

Figure 7: Distribution of oral presentation scores obtained by the case and control groups [Colour figure can be viewed at wileyonlinelibrary.com]

these means (0.25) confirms that the effect size is small, explaining the 90% overlap between both distributions.

This result provides an answer to the second research question (R2). The answer is again nuanced. While there is strong statistical evidence that there is an improvement, the effect size of this improvement is very small (less than 4% of the maximum grade or 8% of the range of grades). Therefore, it can be concluded that the simple addition of automatically generated feedback to learning activities that already provide expert-based feedback has a positive but negligible effect.

Conclusions and recommendations

This work presented an evaluation of the effect of using an automated feedback system to develop oral presentation skills. While there were at least 16 previous evaluations in recent literature, this was the first time that a randomised controlled experiment was performed using one of these systems as part of a real learning task, judged externally by experts and with a student population large enough to draw generalisable conclusions.

The first conclusion of this work is that not all dimensions of oral presentation skills can be improved with just one round of reflective feedback. Looking at the audience during the presentation seems to be very trainable with this automatic system, while the use of filled pauses was barely changed from one session to the other. The main implication of this finding is that instead of focusing on general oral presentation skills, this kind of systems and their evaluation should focus on the best way to train very specific dimensions. For example, training for reducing the use of filled pauses could use interrupting, online feedback, while reflective, offline feedback could be used for maintaining a good posture and looking at the audience. This strategy is already used by human instructors (Ginkel *et al.*, 2015) and should be emulated by automated systems.

The second conclusion is related to the benefits that these systems provide when they are integrated as part of existing learning activities involving oral presentations. Even if a statistically significant positive difference was found, its effect size was small. Given the setting of the evaluation, where all the students received human-generated feedback at some point, this result seems to imply that automated feedback has a lower impact than expert human feedback. The main area of application for the current generation of systems should be courses where there is no planned human-generated feedback.

The present work only evaluated a specific system in a very specific setting with limitations introduced by the requirements of a real learning activity where there is not full control of the experimental design and the particular workings of the systems that restrict the type and variability of the measurements. This first evaluation should not be seen as a definitive answer to the proposed research questions, but as an invitation to other researchers to conduct similar (and better) evaluations to accumulate more evidence on how these systems behave in, and interact with, real learning processes and obtain enough insight to design a new generation of more effective systems to provide improved automated feedback to novice oral presenters.

Statements on open data, ethics and conflicts of interest

The data used for the analysis are available on request from the authors in an anonymised format and within the constraints of Ecuadorian data protection law.

The reported research has been conducted within the ethical regulations in place at the hosting institution of both the researchers, professors and students.

The authors declare no conflicts of interest in the development of this work.

References

Batrinca, L., Stratou, G., Shapiro, A., Morency, L.-P., & Scherer, S. (2013). *Cicero—Towards a multimodal virtual audience platform for public speaking training*. Berlin, Heidelberg: Springer.

Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2017). Realtime multi-person 2D pose estimation using part affinity fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1302–1310). Honolulu, HI. https://doi.org/10.1109/CVPR.2017.143

Chan, V. (2011). Teaching oral communication in undergraduate science: Are we doing enough and doing it right? *Journal of learning design*, *4*(3), 71–79.

Damian, I., Tan, C. S. S., Baur, T., Schöning, J., Luyten, K., & André, E. (2015). Augmenting social interactions: Realtime behavioural feedback using social signal processing techniques. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)* (pp. 565–574). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/2702123.2702314

De Grez, L. (2009). *Optimizing the instructional environment to learn presentation skills*. Belgium: Ghent University.

De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*(2), 385–390.

Fawcett, S. B., & Miller, L. K. (1975). Training public-speaking behaviour: An experimental analysis and validation. *Journal of Applied Behavior Analysis*, *8*(2), 125–35.

Gan, T., Wong, Y., Mandal, B., Chandrasekhar, V., & Kankanhalli, M. S. (2015). *Multi-sensor self-quantification of presentations*. Paper presented at the Proceedings of the 23rd ACM international conference on Multimedia, Brisbane, Australia.

Griffin, P., & Care, E. (2014). *Assessment and teaching of 21st century skills: Methods and approach*. Dordrecht, Netherlands: Springer.

Hamilton, C. (2013). *Communicating for results: A guide for business and the professions*, New York, NY: Cengage Learning.

Kurihara, K., Goto, M., Ogata, J., Matsusaka, Y., & Igarashi, T. (2007). *Presentation sensei: a presentation training system using speech and image processing*. Paper presented at the Proceedings of the 9th international conference on Multimodal interfaces, Nagoya, Aichi, Japan.

Kyllonen, P. C. (2012). *Measurement of 21st century skills within the common core state standards*. Paper presented at the Invitational Research Symposium on Technology Enhanced Assessments, National Harbor, MD.

Miller, J., Lawler-Mcdonough, M., Orcholski, M., Woodward, K., Roth, L., & Mueller, E. (2017). Public Speaking Today. *Stand up, Speak out*.

Ochoa, X., Domínguez, F., Guamán, B., Maya, R., Falcones, G., & Castells, J. (2018). *The rap system: automatic feedback of oral presentation skills using multimodal analysis and low-cost sensors*. Paper presented at the Proceedings of the 8th International Conference on Learning Analytics and Knowledge, Sydney.

Ochoa, X., & Worsley, M. (2016). Augmenting learning analytics with multimodal sensory data. *Journal of Learning Analytics*, *3*(2), 213–219.

Riemer, M. J. (2007). Communication skills for the 21st century engineer. *Global Journal of Engineering Education*, *11*(1), 89–100.

Schneider, J., Börner, D., van Rosmalen, P., & Specht, M. (2015). *Presentation trainer, your public speaking multimodal coach*. Paper presented at the Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, New York.

Schneider, J., Börner, D., Van Rosmalen, P., & Specht, M. (2016). *Enhancing public speaking skills-an evaluation of the Presentation Trainer in the wild*. Paper presented at the European Conference on Technology Enhanced Learning, Cham, Switzerland.

Schneider, J., Börner, D., Van Rosmalen, P., & Specht, M. (2017). *Do you know what your nonverbal behavior communicates?–Studying a self-reflection module for the presentation trainer*. Paper presented at the International Conference on Immersive Learning, Cham, Switzerland.

Schneider, J., Romano, G., & Drachsler, H. (2019). Beyond reality—Extending a presentation trainer with an immersive VR module. *Sensors*, *19*(16), 3457.

Tanveer, M. I., Lin, E., & Hoque, M. E. (2015). *Rhema: A real-time in-situ intelligent interface to help people with public speaking*. Paper presented at the Proceedings of the 20th International Conference on Intelligent User Interfaces, New York.

Tanveer, M. I., Zhao, R., Chen, K., Tiet, Z., & Hoque, M. E. (2016). *Automanner: An automated interface for making public speakers aware of their mannerisms*. Paper presented at the Proceedings of the 21st International Conference on Intelligent User Interfaces, New York.

Trinh, H., Asadi, R., Edge, D., & Bickmore, T. (2017). RoboCOP: A robotic coach for oral presentations. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1*(2), 27.

van Ginkel, S., Gulikers, J., Biemans, H., & Mulder, M. (2015). Towards a set of design principles for developing oral presentation competence: A synthesis of research in higher education. *Educational Research Review*, *14*, 62–80.

Wörtwein, T., Chollet, M., Schauerte, B., Morency, L.-P., Stiefelhagen, R., & Scherer, S. (2015). *Multimodal public speaking performance assessment*. Paper presented at the Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, New York.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.