การพัฒนาเกณฑ์การประเมินทักษะการนำเสนอและความเที่ยงระหว่างผู้ประเมิน สำหรับรายวิชาทักษะการสื่อสารและการนำเสนอเป็นภาษาอังกฤษ

PRESENTATION ASSESSMENT RUBRIC DEVELOPMENT AND INTER-RATER RELIABILITY OF COMMUNICATION AND PRESENTATION SKILLS COURSE

*กรุณา นาผล**

Karuna Naphon*

สถาบันภาษา จุฬาลงกรณ์มหาวิทยาลัย

Chulalongkorn University Language Institute.

***Corresponding author, E-mail:** k.naphon@gmail.com*

บทคัดย่อ

เกณฑ์การประเมินทักษะการนำเสนอสำหรับรายวิชาทักษะการสื่อสารและการนำเสนอเป็นภาษาอังกฤษ (Communication and Presentation Skills: CPS) ถูกพัฒนาขึ้นโดยอาจารย์ผู้สอนในรายวิชาโดยพิจารณาจากเนื้อหาที่สอนในบทเรียนเกี่ยวกับทักษะการนำเสนอและจากประสบการณ์ของอาจารย์ (Intuitively-Based Method) การทบทวนและปรับปรุงเกณฑ์การประเมินดังกล่าวเกิดขึ้นอย่างต่อเนื่องโดยผ่านการประชุมและลงมติร่วมกันของอาจารย์ผู้สอนในที่ประชุมรายวิชาเมื่อสิ้นภาคการศึกษา ค่าทางสถิติของความเที่ยงระหว่างผู้ประเมิน (Inter-Rater Liability) ที่ใช้เกณฑ์การประเมินนี้ถูกทดสอบโดยใช้ Pearson Product-Moment นอกจากนี้ ความคิดเห็นที่ผู้ประเมินเขียนไว้ในแบบฟอร์มเกณฑ์การประเมินสำหรับคู่ของผู้ประเมินที่มีค่าความเที่ยงต่ำถึงปานกลางยังถูกนำมาวิเคราะห์เพิ่มเติมเพื่อหาปัจจัยที่อาจส่งผลต่อความคิดเห็นที่แตกต่างกันของผู้ประเมิน ทั้งนี้ ผลการทดสอบแสดงให้เห็นว่าค่าความเที่ยงระหว่างผู้ประเมินในรายวิชานี้อยู่ในระดับปานกลางถึงสูง อย่างไรก็ตาม การวิเคราะห์ความคิดเห็นของผู้ประเมินทำให้ทราบถึงปัจจัยที่อาจส่งผลต่อค่าความเที่ยง ได้แก่ รูปแบบการให้คะแนนที่แตกต่างกันของผู้ประเมิน เกณฑ์การประเมินบางด้านที่อาศัยความเห็นของผู้ประเมินเป็นหลัก และความผิดพลาดในการให้คะแนนของตัวผู้ประเมินที่เกิดขึ้น ผู้วิจัยได้อภิปรายผลจากการทดสอบที่พบ ข้อแนะนำในการแก้ปัญหาเพื่อปรับปรุงเกณฑ์การประเมินต่อไป ข้อจำกัดของงานวิจัย และแนวทางงานวิจัยในอนาคต

คำสำคัญ: เกณฑ์การประเมินทักษะการนำเสนอ การพัฒนาเกณฑ์การประเมิน ความเที่ยงระหว่างผู้ประเมิน ทักษะการสื่อสารและการนำเสนอเป็นภาษาอังกฤษ

Abstract

A presentation rubric used in a Communication and Presentation Skills (CPS) course at a university in Thailand was developed using intuitively-based method by the course teachers based originally on the course content of the presentation unit and the teachers' experience. It has been revised based on the students' performances perceived by the teachers through the years. The adjustments made were agreed upon at the course meeting at the end of each semester. Inter-rater reliability values were calculated using Pearson product-moment and written comments from the assessors were reviewed to further investigate factors that contributed to low to moderate correlations. The results showed significant correlations at moderate to high levels among the assessors. The written comments revealed some scoring patterns, existing subjective areas, and human errors that should deserve further discussion. A more detailed descriptor of the rubric and teacher training are needed. Pedagogical implications and further research are also suggested.

Keywords: Presentation Rubric Development, Rating Scale Development, Inter-Rater Reliability, Communication and Presentation Skills

Introduction

Related literature, background, and motivation of the study

A performance assessment task, e.g. written compositions, musical performances, or presentations as in this study, requires a student to perform the task while the results cannot be marked using an answer key as for a multiple-choice or true-false test. Performance assessment scoring, therefore, inevitably involves subjective judgments of assessor(s) towards the quality of the student's work. In this case, a good set of rubric is one of the factors that can produce reliable and fair measurement [1-2], help in the process of performance rating, and be a key factor contributing to the reliability of the assessment [3]. Two approaches to rating scale construction categorized by Fulcher, Davidson, and Kemp (2011) [4] are (1) measurement-driven method and (2) performance data-based method. Rubrics from the measurement-driven approach, or intuitive method [5], rely on the experience and knowledge of the designers and can be refined over time based on some theories or experience the designers/users have learned. The scales from this approach may lack description adequacy and may not be contextually specific. On the other hand, rubrics from the performance data-based method, or empirical approach, are derived from a thorough conversation or discourse analysis of the collection of the learners' performance samples. The level descriptors from this method are beneficial for describing the learners' performance in the test task. However, the method is time-consuming and

the raters can find the detailed descriptors difficult to use in real-time rating. Once developed, a rating scale should be trialed and evaluated in terms of its reliability or dependability [6-7]. The consistency between different raters (inter-rater reliability) and the consistency within the same rater in different occasions (intra-rater reliability) are typically tested.

Communication and Presentation Skills (hereinafter CPS) is a speaking and listening course designed for engineering students at a university in Thailand. The four units taught in the course, focusing on socializing, job interview, group meeting and discussion, and presentation, were decided based on a meeting between instructors from the Faculty of Engineering and the language institute who provided the course. The course started in semester two, academic year 2011 and has served year 2-4 students from 10 Engineering Departments. There are 9-11 instructors teaching 10-14 sections in each semester. In this study, the main focus is on the assessment of the solo presentation.

Nine to twelve class hours (in 3-4 weeks) are allocated for the Presentation Skills unit in each semester. These include theoretical and practical aspects as well as a group presentation, which is a classwork worth 5% of the course, basically as a formative assessment task for students to practice giving a presentation before doing a 3-to-4-minute solo presentation in the end. The solo presentation is assessed by two teachers from different sections and accounts for 15% of the course. The students taking the course are informed about the rubric as it is included in the course syllabus and is explained in class to ensure that all students understand the details in the rubric.

The CPS's presentation rubric has been designed and refined based on the measurement-driven approach by the course instructors. Two raters (the course teachers) use the rubric to assess students' presentations. According to Davies et. al. (1999) and Davis & Kendo-Brown (2012) [8-9], a rating scale should consist of three components, which are (1) the lists of criteria or dimension that the performance will be judged, (2) the lists of the scores, and (3) the performance descriptors of each level of the scores. However, the CPS's presentation rubric contains only the lists of the performance criteria and the scores, but does not narrate the performance descriptors of each score level. This type of rubric is called a numerical rating scale [5]. It requires little reading from the assessors when rating but only works when users/assessors have mutual and consistent understanding of what each number means. Case studies and research on measurement-driven or intuitively-driven presentation numerical rating scale as in this study are rare in the literature. However, numerical rating scale is used extensively in speaking and presentation assessments in the researcher's institute, meaning that it may be easy and practical to use and may serve some assessment purposes. The researcher, therefore, aims to illustrate how the presentation rubric has been developed, study whether the rubric developed using

such approach is suitable for the context by investigating inter-rater reliability, and review the written comments of assessors with low to moderate correlations to identify some factors that may contribute to the disagreements.

Objectives

1. To illustrate how the intuitively-based presentation assessment rubric of the course has been developed.

2. To test the inter-rater reliabilities of the pairs of assessors (teachers) in the course.

3. To investigate possible factors contributing to low to moderate correlations.

Significance of the study

The findings will be widely beneficial and crucially important. The course coordinator and teachers will have information to decide what to adjust and/or consider in refining the rubric. Teachers of similar courses may have some guidelines to develop rating scales in their context. The students can be confident that the presentation evaluation of the course is reliable and meaningful to them as what they learn can be roadmap to their skill improvement.

Methods

This section consists of three main parts: the development and revisions of the CPS presentation rubric used in this study; background of the participants; and data collection and analysis.

CPS Presentation Rubric: Its development and revisions

Version 1

Based on the three major components of a presentation – story message; physical message; and visual message [10] – referred to in the course book, together with quality of the content, language use, and time, which are also highlighted in the course, the first version of the solo presentation rubric is shown in Figure 1 below.

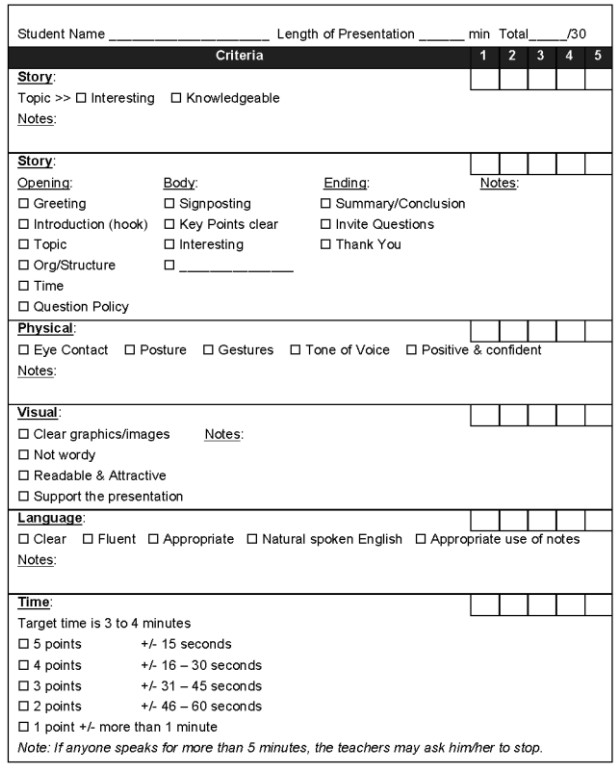
Student's Name:_____ **Length of Presentation** _____**mins** **Total** ____**/30**

Criteria	1	2	3	4	5
Content (interesting, fresh, knowledgeable)					
Organization (opening, signposting, ending)					
Visual aids (in point form, readability, font size, font type, clear graphic images, etc.)					
Body language (gesture, posture, eye contact, clarity of voice)					
Language (clear, fluent, appropriate: uses natural spoken English)					
Time					

Figure 1: Solo presentation rubric_CPS_Version 1

Version 2

In the course meeting at the end of the semester, however, some instructors found it difficult to write comments in such limited space. Moreover, most instructors did not agree on the organization score, especially in terms of the opening; some were satisfied with a basic introduction while some expected a more elaborate or fancier opening. Also, the 'time' was not explicitly defined as to what it meant by 1, 2, 3, 4, or 5 point. Therefore, the rubric was re-designed and made clearer as shown in Figure 2.



Student Name _____ Length of Presentation _____ min Total_____/30

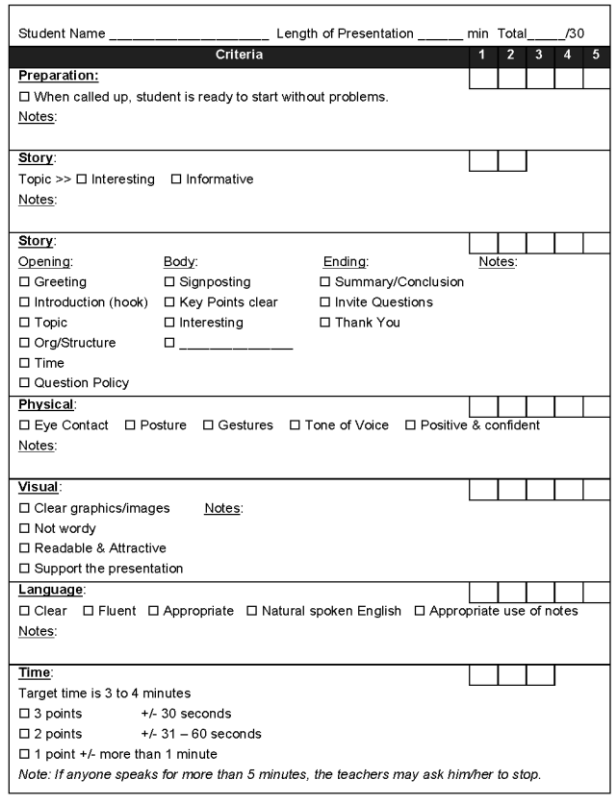
Criteria	1	2	3	4	5
Story: Topic >> ☐ Interesting ☐ Knowledgeable Notes:					
Story:					
Opening: ☐ Greeting ☐ Introduction (hook) ☐ Topic ☐ Org/Structure ☐ Time ☐ Question Policy	Body: ☐ Signposting ☐ Key Points clear ☐ Interesting ☐ _____	Ending: ☐ Summary/Conclusion ☐ Invite Questions ☐ Thank You	Notes:		
Physical: ☐ Eye Contact ☐ Posture ☐ Gestures ☐ Tone of Voice ☐ Positive & confident Notes:					
Visual: ☐ Clear graphics/images ☐ Not wordy ☐ Readable & Attractive ☐ Support the presentation	Notes:				
Language: ☐ Clear ☐ Fluent ☐ Appropriate ☐ Natural spoken English ☐ Appropriate use of notes Notes:					
Time: Target time is 3 to 4 minutes ☐ 5 points +/- 15 seconds ☐ 4 points +/- 16 – 30 seconds ☐ 3 points +/- 31 – 45 seconds ☐ 2 points +/- 46 – 60 seconds ☐ 1 point +/- more than 1 minute *Note: If anyone speaks for more than 5 minutes, the teachers may ask him/her to stop.*					

Figure 2: Solo presentation rubric_CPS_Version 2

Version 3

All the instructors were positive about the rubric version 2 as it was easier for them to check the quality and write comments. This rubric had been used for three semesters (2/2012, 1/2013, and 2/2013) before the second revision. Despite no statistical evidence, a few instructors felt that, on average, the students got higher scores than what they should have got because of the 'time' and 'story (interesting, knowledgeable)' criteria. Some teachers reported that when they listened to a student's presentation, they would have roughly a total score of that student in mind. However, there were some cases when the sum of the scores from all the aspects in the rubric was higher than the total score they initially expected to give the student. For the 'time', those instructors thought that the proportion of 5 points was

too high as many students can get 5 points easily and that increased their total score while their overall quality of the presentation was not that good. The 'story (interesting, knowledgeable)' was found to be quite subjective when assessing as instructors often found the same topic differently interesting. With its 5-point scale, the scores given by two assessors could be quite different. Moreover, most instructors reported that a number of students were not well prepared that they could not start the presentation immediately when called up. Some students had technical difficulties due to lack of preparations. This resulted in time being wasted. Usually, this issue was verbally raised in the class without any reward/punishment. In the 2/2013 course meeting, it was agreed that the rubric be adjusted by reducing the proportion of 'time' to three points and the 'story (interesting, knowledgeable)' that was considered subjective to two points and adding the 'preparation' aspect as it is viewed critical in giving a presentation. Version 3 of the solo presentation rubric of the course shown in Figure 3 below was first used in the first semester of academic year 2014.



Student Name _____ Length of Presentation _____ min Total_____/30

Criteria	1	2	3	4	5
Preparation: ☐ When called up, student is ready to start without problems. Notes:					
Story: Topic >> ☐ Interesting ☐ Informative Notes:					
Story: Opening: ☐ Greeting ☐ Introduction (hook) ☐ Topic ☐ Org/Structure ☐ Time ☐ Question Policy; Body: ☐ Signposting ☐ Key Points clear ☐ Interesting ☐ _____; Ending: ☐ Summary/Conclusion ☐ Invite Questions ☐ Thank You; Notes:					
Physical: ☐ Eye Contact ☐ Posture ☐ Gestures ☐ Tone of Voice ☐ Positive & confident Notes:					
Visual: ☐ Clear graphics/images ☐ Not wordy ☐ Readable & Attractive ☐ Support the presentation Notes:					
Language: ☐ Clear ☐ Fluent ☐ Appropriate ☐ Natural spoken English ☐ Appropriate use of notes Notes:					
Time: Target time is 3 to 4 minutes ☐ 3 points +/- 30 seconds ☐ 2 points +/- 31 – 60 seconds ☐ 1 point +/- more than 1 minute *Note: If anyone speaks for more than 5 minutes, the teachers may ask him/her to stop.*					

Figure 3: Solo presentation rubric_CPS_Version 3

Version 4

At the end of the second semester, academic year 2014, some additional issues were discussed in the course meeting. All the instructors felt that 5 points for 'preparation' was too high, resulting in a higher score than what should have been, considering the overall performance of the students. A few presentations, although instructed clearly, were not engineering-related. Some students were not properly dressed in student uniform. Therefore, 'preparation' in the rubric version 4 was reduced to three points while one more point was added to the 'physical' and 'language' criteria. The requirements regarding the topic and the physical message were stated in the rubric as shown in Figure 4.

Student Name _____ Length of Presentation _____ min Total_____/30

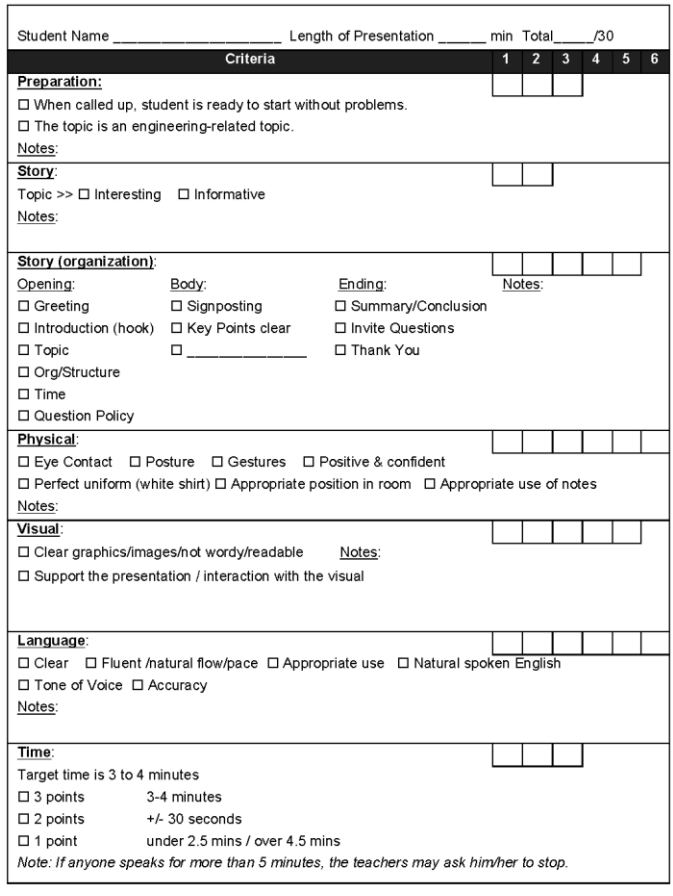
Criteria	1	2	3	4	5	6
Preparation: ☐ When called up, student is ready to start without problems. ☐ The topic is an engineering-related topic. Notes:						
Story: Topic >> ☐ Interesting ☐ Informative Notes:						
Story: Opening: ☐ Greeting ☐ Introduction (hook) ☐ Topic ☐ Org/Structure ☐ Time ☐ Question Policy; Body: ☐ Signposting ☐ Key Points clear ☐ _____; Ending: ☐ Summary/Conclusion ☐ Invite Questions ☐ Thank You; Notes:						
Physical: ☐ Eye Contact ☐ Posture ☐ Gestures ☐ Tone of Voice ☐ Positive & confident ☐ Perfect uniform (white shirt) ☐ Appropriate position in room Notes:						
Visual: ☐ Clear graphics/images ☐ Not wordy ☐ Readable & Attractive ☐ Support the presentation Notes:						
Language: ☐ Clear ☐ Fluent ☐ Appropriate ☐ Natural spoken English ☐ Appropriate use of notes Notes:						
Time: Target time is 3 to 4 minutes ☐ 3 points 3-4 minutes ☐ 2 points +/- 30 seconds ☐ 1 point +/- 31 – 60 seconds *Note: If anyone speaks for more than 5 minutes, the teachers may ask him/her to stop.*						

Figure 4: Solo presentation rubric_CPS_Version 4

Version 5

The rubric version 4 was used in both semesters of academic year 2015. However, interaction with the visual was raised in the meeting as one of the qualities that should be assessed in the course. Tone of voice, which had previously been in the physical aspect, was brought up in the discussion and all the instructors agreed to move it to the language aspect, while 'appropriate use of notes' in the language aspect should be moved to physical message instead. Moreover, an issue related to the 'time' was brought to our attention. One instructor gave a zero to a couple of students whose presentations exceeded five minutes, while his/her assessing partner

gave one point to them. The meeting finally agreed that the students should at least get a point for 'time' if they could stand up and finish their presentation. The rubric was then revised and its version 5, used in semester 1, academic year 2016, which is studied in this research project, is shown in Figure 5 below.



Student Name _____ Length of Presentation _____ min Total_____/30

Criteria	1	2	3	4	5	6
Preparation:						
☐ When called up, student is ready to start without problems.						
☐ The topic is an engineering-related topic.						
Notes:						
Story:						
Topic >> ☐ Interesting ☐ Informative						
Notes:						
Story (organization):						
Opening: Body: Ending: Notes:						
☐ Greeting ☐ Signposting ☐ Summary/Conclusion						
☐ Introduction (hook) ☐ Key Points clear ☐ Invite Questions						
☐ Topic ☐ _____ ☐ Thank You						
☐ Org/Structure						
☐ Time						
☐ Question Policy						
Physical:						
☐ Eye Contact ☐ Posture ☐ Gestures ☐ Positive & confident						
☐ Perfect uniform (white shirt) ☐ Appropriate position in room ☐ Appropriate use of notes						
Notes:						
Visual:						
☐ Clear graphics/images/not wordy/readable Notes:						
☐ Support the presentation / interaction with the visual						
Language:						
☐ Clear ☐ Fluent /natural flow/pace ☐ Appropriate use ☐ Natural spoken English						
☐ Tone of Voice ☐ Accuracy						
Notes:						
Time:						
Target time is 3 to 4 minutes						
☐ 3 points 3-4 minutes						
☐ 2 points +/- 30 seconds						
☐ 1 point under 2.5 mins / over 4.5 mins						
Note: If anyone speaks for more than 5 minutes, the teachers may ask him/her to stop.						

Figure 5: Solo presentation rubric_CPS_Version 5

A series of adjustments has been made following the teachers' suggestions, with quite a few major details revised, including the criteria and the scores.

There was, however, no concern regarding the solo presentation rubric raised in the course meeting at the end of semester 1/2016. The researcher, therefore, would like to study whether there is significant correlation between the two raters when assessing the solo presentation using this rubric. Written comments by the pair(s) of assessors whose correlation is low are also reviewed.

Participants of the study

Participants of the study were 10 teachers, eight native speakers and two Thai teachers, teaching 324 students (11 sections with 27-31 students each) in semester 1/2016. All the teachers consented to participate in

the study. The solo presentation schedule of the last two weeks of the semester together with the assessors and background information of the teachers are as follows:

Presentation Week 1		
Time slot	Sections	Assessors
Day 1	1	B & C
	2	A & D
Day 2	5	F & H
	6	E & G
Day 3	9	C & J

Presentation Week 2		
Time slot	Sections	Assessors
Day 1	3	B & D
	4	A & C
Day 2	7	E & H
	8	F & G
Day 3	10	C & I
	11	H & J

Figure 6: CPS_Solo presentation schedule (1/2016)

Remarks: The section numbers and teachers' name were changed in this paper.

Teachers	Own section(s)	Gender	Native / Thai	Started teaching CPS	Teaching other presentation courses?
A	1	M	Native	2/2013	Yes
B	2	M	Native	1/2012	Yes
C	3, 11	M	Native	1/2012	Yes
D	4	F	Thai	2/2012	Yes
E	5	M	Native	2/2014	Yes
F	6	M	Native	2/2014	Yes
G	4	M	Native	1/2012	Yes
H	8	F	Thai	2/2011	Yes
I	9	M	Native	2/2014	Yes
J	10	F	Native	2/2011	Yes

Figure 7: CPS_Teacher information

From Figure 7, only Teacher C taught two sections in semester 1/2016. There are three female teachers and two Thai teachers. All the teachers have more than five years of experience in teaching English and at least two years of experience in teaching CPS and they all teach other presentation courses as well.

Data collection and analysis

The scores given by the two assessors of all the 11 sections were input in the Statistical Package for the Social Sciences (SPSS-version 22) program and were analyzed for inter-rater reliability using the Pearson product-moment (*r*). Based on Brown (2005) [11], correlation coefficients of .00 to .59 are considered to be low, while correlations of .60 to .79 can be viewed as moderate, and correlations of .80 to 1.00 are considered high. In this study, the correlation of the total scores of any pair of assessors that was less than .80 (high correlations) was investigated further by calculating the Pearson *r* for each presentation criteria and reviewing the assessors' written comments to see which aspects caused such differences.

Results

For the second research objective, Pearson correlation coefficients (*r*) of each pair of assessors are shown in Figure 8 below.

Teachers	A	B	C	D	E	F	G	H	I	J
A	1.00		.74** (Sec 4 N=30)	.66** (Sec 2 N=31)						
B		1.00	.88** (Sec 1 N=27)	.87** (Sec 3 N=29)						
C			1.00						.71** (Sec 10 N=30)	.71** (Sec 9 N-29)
D				1.00						
E					1.00		.90** (Sec 6 N=31)	.84** (Sec 7 N=30)		
F						1.00	.67** (Sec 8 N=29)	.84** (Sec 5 N=29)		
G							1.00			
H								1.00		.96** (Sec 11 N=29)
I									1.00	
J										1.00

** Correlation is significant at $p < .01$ level (2-tailed).

Figure 8: Matrix of correlation coefficients (r) of all the 11 pairs of assessors in 1/2016

From the figure above, sections 1, 3, 5, 6, 7, and 11 (55%) are highly correlated ($r > .80$, $p < .01$), while sections 2, 4, 8, 9, and 10 (45%) are moderately correlated ($.60 < r < .79$, $p < .01$), and none of them are low correlations. Although the *r* values are not low, they are worth investigating to find out which aspects may have been responsible for the disagreements between the assessors. Therefore, the Pearson *r* for each of the aspects from the presentation rubric for sections 2, 4, 8, 9, and 10 were calculated and shown below.

Criteria / Sections	Section 2 A & D	Section 4 A & C	Section 8 F & G	Section 9 C & J	Section 10 C & I
Preparation (3 points)	.78**	.30	-.04	.68**	.93**
Story (2 points)	.15	.24	-.09	.40*	-.05
Organization (5 points)	-.02	.24	.48**	.58**	.09
Physical (6 points)	.41*	.55**	.57**	.31	.22
Visual (5 points)	.23	.37*	-.06	.38*	.32
Language (6 points)	.09	.55**	.60**	.52**	.76**
Time (3 points)	1.0**	.73**	1.0**	.89**	.94**

* Correlation is significant at $p < .05$ level (2-tailed).

** Correlation is significant at $p < .01$ level (2-tailed).

Figure 9: Pearson r values for each of the presentation aspects of the sections whose overall r < .80 ($p < .01$)

From Figure 9 above, Pearson *r* correlations of each aspect of each pair range from minus values to perfect 1.00-correlation values. Some correlations are statistically significant, while others occur by chance. Overall, among these five pairs of assessors, Story, Organization, and Visual aspects appear more problematic than others as most of the correlation values are not statistically significant. In order to see what may be the factors contributing to these discrepancies, written comments from the assessors of each aspect of each section with low correlation coefficients ($r < .60$) were all reviewed. Some major findings are described below.

For 'Preparation', two out of 30 students in section 4 got full scores from Teacher A but got only 1 point from Teacher C with '*No preparation, downloaded (slides) from internet and came in late*' comments. Both teachers gave equal preparation scores to all other students in the section. In section 8, one student got 3 points from Teacher F but got 1 point from Teacher G. No specific comment was written for 'preparation' from Teacher G, but overall comment was '*not impressive*' and that seemed to lower the scores in all aspects for this student.

'Story' seems to be problematic considering the Pearson *r* in the table above. From reviewing the score sheets,

most teachers did not write much on this. No particular pattern was found, but Teacher C gave 1.5 to a number of students while other teachers did not give a half point.

In terms of 'Organization', the differences were 1 to 2 points for section 2. Teacher A and D marked down the scores because of '*slow developing hook*' and '*lack conclusion*.' For section 4, Teacher A's major comments were also for abrupt conclusion. Teacher C ticked the boxes of the organization components clearly. Five points was given to students who covered all the points, while 4.5 and a few 4 were given to those who missed a few points or covered the points but not with good quality. For section 8, the differences in 'organization' scores were 1 point and Teacher G was usually more lenient than Teacher F. Apart from the tick marks, major negative/positive comments from both Teachers were on 'hook' in the introduction. For section 9, the teachers deducted 'organization' points mostly because of unnecessarily long introductions and misuse of signposts/lack of organization of the body part. For section 10, the differences ranged from 0.5 to 3 points. From reviewing the written comments, there were some students that one teacher gave a perfect organization score with '*pretty good / good try for the hook*' comments, while the other teacher gave only 2 or 3 points with comments such as '*unclear key points / flat story / don't get the hook*.'

Regarding the 'Physical' aspect, the differences between the assessors for sections 2, 4, and 8 were mostly 0.5 to 1 point and 2 points for only a few students. Most comments showed that the assessors for section 2 deducted the points because of ineffective gestures, whereas the assessors for section 4 gave comments on weak eye contact (looking at computer screen) and ineffective walking movements when the students walked and stood in front of the screen. Most positive comments from the assessors for section 8 were for confidence and enthusiasm of the students, while the negative comments were on the students' posture of being stiff or stagnant. For sections 9 and 10, it clearly showed that Teacher C was strict on 'Appropriate position in room' as '*student standing in light / behind the computer*' were major comments written for students with low 'physical' scores. On the other hand, Teachers J and I focused on lack of eye contact, gestures, or confidence of the speakers. This caused the differences as high as 2 to 2.5 points for a few students in these sections.

The overall score differences of 'Visuals' in these five sections were 0.5 to 1.5 points. Apart from general format comments such as '*wordy slides, small texts, script on, grammar/spelling mistakes*,' a number of 'subjective' comments were found. For example, '*not very exciting slide, boring template, mixed quality (some good, some dull/weak slides), not support the presentation much, very nice slides, low color contrast, excellent, good use of video*.' Different scoring patterns are also noted. Teachers A generally gave

4 points (no written comment) and gave 3 points when some negative qualities were noted. Teachers D, C, F, G, J, and I generally gave 5 points to students with no particular quality noted, gave 4 points to students with one negative quality, and gave 3 points (or 3.5 points for Teacher C) to students with two and more negative qualities. Some major discrepancies are found in sections 8, 9, and 10. In section 8, there are four students with 2-point difference in visual scores. They got the perfect visuals score from Teacher G but got only 3 points from Teacher F with '*mixed quality, could be more interesting, dull*' comments. For section 9, one student got '*mostly very good*' (5 points) from Teacher C but got 3 points with '*spelling, busy pictures, unreadable map*' comments from Teacher J. In section 10, one student got 5 points from Teacher I saying '*nice visuals*' but got 3 points from Teacher C with '*grammar / lots of writing*' comments.

In terms of 'Language,' the assessors for section 2 did not seem to agree at all. The score differences were between 1 to 2 points. Teacher A generally gave 5 points to students with '*generally ok, smooth delivery, and natural language*,' gave 4 points to some '*rambling and fast speaking*,' and gave 3 points to a few '*slow delivery*' presentations. Teacher D gave a perfect score of 6 points to two students (no written comment), gave 5 points to students with '*fluent and clear*' language, gave 4 to '*trembling voices*,' gave 3 to students with two or more points of '*not good flow, chunk, fast, no intonation, script remembered, no ending sound in pronunciation*,' and gave 2 to students who '*can't remember the script, reads the script, speaks with unclear pronunciation* (sound level).' However, two students got 3-point difference in their scores. One student got 5 points from Teacher A ('*pretty smooth speaking*' comment), but got only 2 points from Teacher D ('*unclear pronunciation of many words*' (sound level) comment). The other student got 3 points ('*slow delivery*' comment) from Teacher A, but got 6 points from Teacher D (no comment). For sections 4 and 9, the score differences are 0.5 to 1 points. Teacher A gave 6 points to a few students with '*very well rehearsed*' comment, gave 5 to students with '*not very natural but well-memorized, loud and clear, good energy*' comments, gave 4 to '*slow delivery, gets stuck several times, not very natural*' presentations, and gave 3 to students with '*halting / rough delivery, many errors*' when speaking. Teacher C is the only one who gave a half point. For example, 5.5 points for '*natural, pretty good for all the criteria*,' 4.5 to 5 points for '*sluggish delivery*,' 4 points to students who '*speak very quickly, choppy, robotic*,' 3.5 points for '*too flat & hesitant, no energy*'; and 3 points for '*Thai pronunciation and monotone*' comments. Teacher J generally gave 5 points for '*clear but some pronunciation problems*'; 4 points for '*erratic, not natural flow, grammar mistakes, problems in 'ed' endings, slow pace, soft voice*'; 3 points for students with '*weak language structure and pronunciation problems*.'

Although the correlation coefficients of 'Time' were all moderate to high (r = .70 – 1.00), it is the only objective criteria that should have a perfect correlation of 1.00, as whether the students finish the presentation in time can be measured clearly. However, three pairs of teachers are found to have disagreement on this point. In section 4, there were five students with one point difference in 'Time.' Among these, three of them seemed to be from that a teacher was not very strict on the time. Teacher C wrote 2:54, 2:55, and 2:54 minutes while Teacher A wrote 3 minutes for all of them. The other two students were obviously from human errors as Teacher A wrote 2:37 and 4:20 minutes but gave the full score of three points, instead of two points, to the students. In section 9, there were three students with one point difference in 'Time.' Again, Teacher C was strict on the time whereas Teacher J compromised on this as Teacher J wrote 2:56 minutes for two students but gave the full score of three points to them. The difference for another student was from human error as Teacher J wrote 4:42 minutes but gave two points, instead of one, to the student. The same situations happened for section 10. Teacher C was strict on the time and correct when interpreting the time to the score while Teacher I compromised the criteria for one student and wrongly interpreted the time to the score for one student.

Conclusions and Discussion

The presentation rubric development of the course

The CPS presentation rubric was an analytic rubric developed and revised based on the measurement-driven method by the course teachers. Some more points of concern have been added and the weight of the scores has been adjusted, based on the students' performances that the teachers have witnessed through the years. From the revising history of the rubric described earlier, the weight of the scores, however, seems to be a result of a holistic judgment of the overall performance, not an analytic assessment. The teachers 'felt' that with an equal weight, some subjective aspects can make the students get a 'too high' score. This is in line with Davis and Kondo-Brown (2012) [8] who pointed out that, in practice, there are times when teachers score holistically even when using an analytic rubric.

From reviewing the written comments, some improvements are needed for this numerical rating rubric itself as well as for the better consistency of the teachers' judgment. For example, more detailed descriptors may be needed, e.g. for each level of the language, or counting rules need to be clarified if, for example, components of the Organization are more practical to count than writing up new descriptors for this aspect. However, if the quality of the component is another factor, the checked box counting, though making the scoring more objective, is not the ideal answer [12].

One of the strengths of the course is that the teachers are all very well cooperative and critical in giving feedback to anything they see that needs attention. All the adjustments made to the rubric are from a thorough discussion among more than 80 percent of the course teachers who attended the course meeting at the end of each semester. The current version of the rubric consists of the criteria in line with 10 points out of the 11 core and optional performance standards for the Public Speaking Competence Rubric (PSCR) by Schreiber et al. (2012) [13] (The 11th point is the optional standard for a persuasive speech which is not the focus of this course.) So far, the current rubric, together with specific written comments from the teachers, can be considered useful and suitable for the context as it is easy to use in real-time rating and it can provide feedback to the students [8]. Nonetheless, it is not the best version. Revisions of a rubric should still be done from time to time when necessary [14].

Inter-rater reliability of the scores given by pairs of the course teachers

Of all the 11 sections, the Pearson *r* correlations show the moderate to high inter-rater reliability ($r > .60$, $p < .01$) among the assessors. The written comments of five pairs that are moderately correlated were reviewed and some points of concern/ patterns were found.

First, scoring patterns of some teachers are noted. As described above, Teacher G may sometimes consider the holistic picture rather than an analytic one as he seemed to base his score for a student in section 8 on his overall impression. Interestingly, Teacher C is the only one who gave a half point. This numerical rating rubric does not have a clear descriptor of each score point, so it is worth investigating more with Teacher C how he came up with a half point. Also, this can lead to necessity in having clear descriptors for each level or having clear rules of how to give the score if the teachers think 'counting' the checked box is more practical than writing up and using detailed descriptors. When no specific comment is noted, some teachers generally gave a full point, while some generally spare one point for a really outstanding performance. Again, clear rules of how to score or how to count the checked box may be needed.

Second, some subjective areas contributed to the score differences. These include the 'Story,' the 'Hook' of the introduction part, and whether the 'Visual' (Power Point slides) is good or dull. As described above, a number of 'subjective' words are noted, e.g. not exciting slides, flat hook, nice/dull/weak slides. In this case, teacher training can play a role in shaping the teachers' perspectives.

Moreover, some negative qualities seem to bother one assessor more than others. For example, '*standing in light*' was frequently noted by Teacher C and pronunciation (segmental level) was noted more often than others by Teacher D, who is one of the two Thai teachers in the course.

Zhang and Elder (2011) [15] also noted a similar point that while comments on 'Language' from native speakers concern more on intelligibility of the presentation, a non-native one seems to pay attention to specific sounds.

Finally, human errors occurred. This is obvious especially for the 'Time' aspect as described in the result section above. Even for a very objective aspect like this, the assessors still showed unreliable decisions when one was more compromising than the other, or when the interpretation of the time to the score was wrong. Apart from that, some written comments may suggest a human error causing the difference as well, for example, when the teachers did not agree on whether the student was late or unprepared (downloading the slide in class), or whether the slides were 'wordy' or not. These negative qualities should have been obvious for both assessors to see. In these cases, rater training and discussion are crucial keys to mutual understanding among the assessors [16-18].

Implications

Although some scholars [4] noted a few drawbacks of the approach as mentioned earlier, from the moderate to high correlation values, it can be implied that the intuitively driven CPS presentation rubric by the course teachers is suitable for the context. However, it is necessary that descriptors for each level of the score or other scoring rules and teacher training be done to reduce rating variability [16-18]. Until the calibration is ready, based on this study, the teachers should have a discussion on the score when there is a difference of the total score of more than 3 points, as it is found that the score differences of the six sections with high correlation values are between 0 to 3 points, and wider gaps are found in all the five sections with moderate correlations. Pedagogically, teachers may use the rubric to train the students for self- or peer-evaluations so that they can develop autonomous learning as well [13].

Limitations and future studies

The data for the inter-rater reliability testing from just one semester in this study provides a small sample size. More studies with bigger sample size can be done by testing the presentation scores across semesters/years of the pairs of assessors that work together. Also, the qualitative data (discussion with teachers and written comments) was less explored as it is not the main focus of this study. Future studies may include an in-depth interview with the teachers to see their perspectives of how they use the rubric. Some other characteristics of the assessors may also be investigated, such as, their experience (the years of teaching or how many presentation courses they teach), and whether they are native or non-native speakers of English. In addition, research on whether the rubric helps the students in learning or preparing for the presentation is worth studying.

Conclusion

The inter-rater reliability of the intuitively-driven numerical rating rubric for a presentation task was moderate to

high. This suggests that the development of rating scales using intuitively-driven approach, with constant review, can be a suitable practice for the context. However, the written comments showed some existing subjective aspects that need to be further discussed among the course teachers. In addition, descriptors of each level of the score should be narrated in order to increase mutual and consistent understanding of the teachers and decrease the score variability. Further studies should be done with a bigger sample size and on other characteristics or related conditions, for example, the assessors' characteristics or the students' perception towards the rubric as a means to help with their learning.

References

[1] Tim McNamara. (1996). *Measuring second language performance.* London & New York: Longman, 117-129.

[2] Perlman, Carole. (2012). An introduction to assessment performance scoring rubric. In Carol Boston (Ed.), *Understanding scoring rubrics: A guide for teachers* (pp. 5-13). Retrieved March 5, 2017, from https://eric.ed.gov/?id=ED471518

[3] Turner, Carolyn E. (2013). Rating scales for language tests. In Carol A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1-7). Chichester, West Sussex, UK: Wiley-Blackwell.

[4] Fulcher, Glenn; Davidson, Fred; and Kemp, Jenny. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing.* 28(1): 5-29. doi:10.1177/0265532209359514

[5] Sari Luoma. (2004). *Assessing Speaking*. Cambridge: Cambridge University Press.

[6] Lyle F. Bachman; and Adrian S. Palmer. (1996). *Language testing in practice: Designing and developing useful language tests.* Oxford: Oxford University Press.

[7] H. Douglas Brown. (2004). *Language assessment: Principles and classroom practices.* White Plains, NY: Pearson Longman.

[8] Davis, Larry; and Kondo-Brown, Kimi. (2012). Assessing student language performance: Types and uses of rubrics. In J.D. Brown (Ed.), *Developing, using, and analyzing rubrics in language assessment with case studies in Asian and Pacific languages* (pp. 33-55). Honolulu: University of Hawai'i, National Foreign Language Resource Center.

[9] Davies, Alan; et. al. (1999). *Dictionary of language testing*. Cambridge: Cambridge University Press.

[10] David Harrington; and Charles LeBeau. (1996). *Speaking of Speech.* Tokyo: MacMillan Language house.

[11] James Dean Brown. (2005). *Testing in Language Programs: A comprehensive guide to English language assessment.* New York: McGraw-Hill.

[12] Nathan T. Carr. (2011). *Designing and Analyzing Language Tests.* Oxford: Oxford University Press.

[13] Schreiber, Lisa M.; Paul, Gregory. D.; and Shibley, Lisa R. (2012). The Development and Test of the Public Speaking Competence Rubric. *Communication Education.* 61(3): 205-233.

[14] Mertler, Craig A. (2012). Designing scoring rubric for your classroom. In Carol Boston (Ed.), *Understanding scoring rubrics: A guide for teachers* (pp. 72-81). Retrieved March 5, 2017, from https://eric.ed.gov/?id=ED471518

[15] Zhang, Ying; and Elder, Catherine. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs?. *Language Testing*, 28(1): 31-50.

[16] Davis, Larry. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*. 33(1): 117-135. doi:10.1177/0265532215582282

[17] Fulcher, Glenn. (2015). Assessing second language speaking. *Language Teaching.* 48: 198-216. doi:10.1017/S0261444814000391

[18] Joe, Jilliam; et al. (2015). A prototype public speaking skills assessment: An evaluation of human scoring quality (ETS RR-15-36). *ETS Research Report Series*. doi:10.1002/ets2.12083