Presentation Peer Assessment: Friendship Matters?

Robert Vaughan
Rikkyo University
Yukie Saito
Rikkyo University
Yukie Saito
Waseda University

Reference Data:
Vaughan, R., Saito, Y., & Saito, Y. (2016). Presentation peer assessment: Friendship matters? In P. Clements, A. Krause, & H. Brown (Eds.), *Focus on the learner*. Tokyo: JALT.

Presentations are widely used in EFL classrooms. Peer assessment (PA) of presentations is sometimes employed to promote learner autonomy, listening skills, and metacognitive knowledge. Despite the benefits of presentation activities, some challenges remain. For example, presentations can be stressful, especially for introverted learners, and learners' success in presentations might be determined by possession of advantageous personality traits. Moreover, PA could be biased due to the degree of friendship perceived to exist between the assessor and the assessed. We investigated 3 issues with 117 university students: the correlation between presentation scores and personality, the reliability of PA compared with teacher assessment (TA), and the influence of students' friendship on PA. Results showed that students' personalities did not significantly affect their presentation scores. Moreover, they evaluated peers' presentations in a generally unbiased manner, which indicated that they understood the constructs, although there was still a discrepancy between PA and TA.

プレゼンテーション活動には様々な利点があるが、問題点も残る。例えば、プレゼンテーションは、内向的な学習者にとってストレスとなる可能性があり、外向的な学習者にとって有利になる可能性がある。もう一つの懸念は、学習者の自律性、リスニングスキル、メタ認知知識を伸ばすためにピア評価がよく取り入れられるが、評価する側とされる側の友人関係の度合が、ピア評価に影響を及ぼすという点である。これらの問題をふまえ、本研究では117人の大学生を対象にプレゼンテーションと彼らの性格との関連性、ピア評価と教師評価との信頼性、学生間の友人関係がピア評価に与える影響の3つの点を調査した。その結果、学生の性格はプレゼンテーションに影響を与えておらず、学生は友人関係に左右されずにピア評価を行っていることが判明した。このことは学生がピア評価項目を理解していることを示しているが、ピア評価と教師評価には差異があることも示唆している。

Speeches and presentations are often used as activities in EFL classrooms from junior high school through university, both in courses that focus specifically on presentation skills and also as part of other communication classes. The perceived benefits of presentations include that they allow students opportunities for output, extensive speaking, and fluency work. Classes with speech and presentation activities often adopt a process approach in which learners revise their presentations several times before the final performance and receive feedback from teachers and peers. Peer assessment (PA) often forms part of this process, not only because it provides richer feedback, but also because it develops students' knowledge and use of metalanguage, thus promoting self-regulated learning (De Grez, Valcke, & Roozen, 2012).

Despite the positive aspects of presentation activities, some reservations remain about their use in EFL learning. Presentations, which usually involve speaking in front of the class, have been found to be related to high foreign language anxiety (Mak, 2011). Besides anxiety, successful classroom performance including presentation may also be linked to personality traits such as confidence or extroversion (Rothstein, Paunonen, Rush, & King, 1994). There appears to be a connection between personality and foreign language classroom anxiety (Apple, 2011; Dewaele, 2013), yet to our knowledge, few previous studies have investigated the direct link between personality and presentations.

Another reservation is the reliability of PA. Studies that have compared teacher assessment (TA) with PA (Patri, 2002; Saito, 2008) have indicated that training sessions in which students practice using an evaluation rubric can enhance the reliability of PA, yet students might interpret the rubric differently from the teacher (Patri, 2002). It should be

noted that these studies treated the teacher as an expert rater, though empirical research, particularly in the field of testing, has repeatedly shown that teachers themselves are not completely free from bias (e.g., Lumley & McNamara, 1995). One possible reason for PA being less reliable could be a bias induced by the degree of closeness or friendship perceived to exist between the assessor and the assessed. Cheng and Warren (1997), who investigated university students' attitudes toward PA in Taiwan, found that their limited English proficiency made students feel unqualified to evaluate peers' performance. Moreover, some students reported their tendency to award a higher score to classmates with whom they were friendlier.

In this study our goal was to investigate the three possible adverse factors we mentioned earlier that could affect the grading of presentations: (a) the relationship between students' personality traits and the grades they received, (b) differences in students' and teachers' understanding of the grading, and (c) bias in peer assessment as a result of level of friendship.

Objectives

We had three objectives. Our first objective was to investigate whether there was a correlation between TA of students' presentations and items on the Japanese Big-Five scale of personality traits (Namikawa et al., 2012). With some evidence that learners feel anxious in presenting in front of other students (Mak, 2011) and that personality is related to speaking anxiety (Apple, 2011; Dewaele, 2013), it was hypothesized that personality traits might affect presentation performances, and thus that categories of the short form of the Japanese Big-Five scale might correlate with presentation performance assessments. The second objective was to test the reliability of PA compared with TA. Based on previous findings (Patri, 2002; Saito, 2008), we hypothesized that there would be modest reliability between PA and TA because, although it was expected that the assessment instrument would bring PA and TA closer together, other factors might have an effect such as differences in experience. The third objective was to examine the influence of students' social networks on PA. As was suggested in previous research (Cheng & Warren, 1997), we hypothesized that biased scores in PA would be related to the degree of closeness between the assessor and the assessed.

Method

The Context

The participants were 117 university students from a large private university in Japan, drawn from three concurrent EFL classes titled *Travel and Culture* taught by the same teacher. Students were asked to give a presentation describing a place or an aspect of culture. Before the presentations took place, the participants were asked to fill in questionnaires on two occasions. The questionnaires were (a) a form developed by the authors in order to ask participants about social networks in the classes (see Appendix A) and (b) the short form of the Japanese Big-Five scale of personality traits (Namikawa et al., 2012; see Appendix B). The presentations occurred near the end of the course and scores represented 20% of students' final grades. They were conducted in four groups of 10 or 11. Each student in the group presented once and assessed the other group members. The teacher observed and assessed all the presentations. After consulting literature on the topic (Yamashiro, 1999) and EFL presentation textbooks (Gershon, 2008; Harrington & LeBeau, 2008), we decided to measure five aspects of presentations: (a) time; (b) memory / looking at notes; (c) speaking / physical performance; (d) story / content; and (e) visuals (poster).

Peer-Assessed Presentations

The peer assessment form and the presentation were designed to achieve several pedagogical goals: to provide reliable and valid evaluation of the skills represented in the constructs, to develop and assess the skills learned throughout the semester, to provide feedback for learners and develop their oral communication skills, to prompt reflection on performance, and to develop learners' listening skills.

The assessment form (see Appendix C) was developed with a rating scale of 1-5 for each construct. Each rating level is described in order to make the form easy to use, to clearly describe different levels of performance, to allow learners to assess constructs they understand, and to provide feedback to learners. The reliability of PA and the assessment instrument were investigated through the analyses described in the following section. As time was a purely objective measurement, it was not included in our analyses.

Analyses

In order to examine the three hypotheses presented earlier, we conducted a series of statistical analyses. The analysis for the first hypothesis (i.e., the relationship between personality traits and presentation performance) was twofold: confirmatory factor anal-

ysis (CFA) and covariance structure analysis (CSA). CFA is used to test a hypothesis about the relationships between latent variables, that is, variables that are believed to exist that cannot be measured (Field, 2009). In our study, the latent variables were the Big-Five personality traits (neuroticism, extroversion, openness, agreeableness, and conscientiousness). We used CFA to see if the first four items from each of the five traits from the Japanese Big-Five scale of personality traits (see Appendix B) would be valid. Based on the results of the CFA, we then conducted the CSA, which is employed to make estimates of the extent and significance of cause and effect relations between variables that researchers hypothesize (Garson, 2008). In this case, we examined the hypothesized effect of different personality traits on presentation performance using AMOS, a software program for CSA and structural equation modeling (Toyota, 2007). To check the model fit, the comparative fit index (CFI) and the root mean square error of approximation (RMSEA) were used to evaluate the model fit. A CFI value above 0.9 and a RMSEA value below 0.1 indicate acceptable model fit (Toyota, 2007).

Regarding the second hypothesis (i.e., the correlation between PA and TA), Cronbach's alpha analysis was conducted using SPSS. The scores students received from peers were averaged and used as PA and were compared with those the teacher gave (TA). Additionally, percent agreement rates between PA and TA in identifying high-rated and low-rated presentations were also calculated. Some methodologists (e.g., Krippendorff, 2004) are against the use of percent agreement because it could hide important disagreement and could be vulnerable to gaming (see Joyce, 2013, for further discussion), but when used as a complement to other reliability indices, it can offer meaningful results (Lombard, Snyder-Duch, & Bracken, 2010).

As for the third hypothesis (i.e., the effects of social networks on PA), we conducted a Rasch analysis, which is used to test how severe or lenient judges of performance are and thereby indicate bias (Bond & Fox, 2015). We employed the analysis to test for potential bias in the ratings students gave each other. Following previous Rasch research (Lumley & McNamara, 1995) the infit mean square values were checked for each rater. A general rule for acceptable values is the range from 0.7 or 0.8 to 1.2 or 1.3 for well-balanced data, but this is not suitable for messy ratings (Bonk & Ockey, 2003). In our research, the acceptable value was set from 0.7 to 1.5. Finally, we conducted a two-way contingency analysis, which is used to check if there is a significant interaction between categorical variables. In this study, the variables were the presence or absence of bias and the degree of friendship.

Results

Hypothesis 1

The results of the CFA yielded a moderate fit, indicating that the 20 personality test items selected for this study were valid measures of the five personality traits ($\chi^2 = 282.48$, $df = 160$, $p < .001$, CFI = .77, RMSEA = .08, AIC = 422.48). Although the resultant CFI of .77, did not meet the criteria, the RMSEA was below.1, which we considered a sign of moderate fit. All of the *p* values were below .05, which shows they were significant; therefore, we decided to conduct a CSA (see Appendix D for the AMOS output).

According to the CSA, the result of the model fit was CMNI = 467.53, $df = 247$, CFI = .71, RMSEA = .09. Again, two indicators were used to evaluate a model fit. The CFI was below the benchmark of .9 and RMSEA was just below 1.0, so we considered the model to have an acceptable fit and continued our analyses. None of the paths from the five factors to the presentation scores were significant. Table 1 shows the result of the CSA and Appendix E shows the AMOS output. Therefore, we concluded that in contrast to our hypothesis, there was no correlation between teacher-assessed presentation performance and the personality traits as measured by the Big-Five scale of personality traits.

Table 1. The Results of the CSA Between Presentation and Personality Scores

Item			Standardized regression	*p* value
Presentation	<---	Neuroticism	.101	.221
Presentation	<---	Extraversion	.111	.396
Presentation	<---	Openness	.477	.065
Presentation	<---	Agreeableness	-.221	.118
Presentation	<---	Conscientiousness	-.113	.118

Note. CSA = Covariance structure analysis

Hypothesis 2

The correlation coefficient between PA and TA for the total score was .48, below the suggested benchmark of .70 (Larson-Hall, 2009), indicating that PA and TA did not agree in their measurements. However, when we looked at the correlation coefficient for the four components of presentations separately, there was a modest correlation found only with

visuals (α =.74). One reason for this could be that the assessment criterion for visuals was easier for student raters to understand. They were asked to assess the quality of visuals and also the degree to which presenters used them effectively to support their ideas. Compared to visuals, the criteria for the other three presentation components (memory, speaking, and story) might have been more ambiguous, resulting in lower reliability between PA and TA (α = .33, .47, and .26, respectively).

Table 2. Summary of Inter-Rater Reliability Analysis

Item	TA (mean)	PA (mean)	Cronbach's α
Memory	3.08	3.62	.33
Speaking	3.75	4.11	.47
Story	4.15	4.23	.26
Visuals	3.54	4.03	.74

Note. TA = teacher assessment; PA = peer assessment.

The results of the reliability analysis with Cronbach's α showed that there was relatively low agreement between PA and TA, which is consistent with the findings of De Grez et al. (2012). The question that appeared to be more important pedagogically, however, was whether students could distinguish high performers from low performers who failed to meet the criteria. For this purpose, percent agreement rates between PA and TA in identifying the high-rated and low-rated presentations were also calculated.

First, out of the 117 presentations, 25 high-rated presentations, which scored in the high range of 18 to 20 in TA, and 27 low-rated presentations, which scored in the low range below 11 in TA, were identified (see Table 3). Only five of 25 high-rated presentations were identified as such by both TA and PA, resulting in a low agreement rate (20%). Regarding the 27 presentations with the lowest ratings, 18 were identified as such by both TA and PA, leading to a higher agreement rate between TA and PA (66.7%). The score range for the high-rated presentations (18-20) was much smaller than that for the low-rated presentations (5-11), which could possibly explain why raters were more successful in identifying poor performers than high performers.

Table 3. Summary of Percentage Agreement Between TA and PA Scores

Presentations	Score range in TA	Number of presentations		Agreement (%)
		Identified in TA	Identified in PA	
High-rated	18-20	25	5	20.0%
Low-rated	5-11	27	18	66.7%

Note. TA = teacher assessment; PA = peer assessment.

Hypothesis 3

To investigate the hypothesis that biased scores in PA would be related to the degree of closeness (i.e., whether or not the rater and rated student were friends), a facets analysis of PA was conducted with three facets: presenters (n = 117), raters (n = 117), and items (memory, speaking, story, and visuals). Memory turned out to be the most difficult criterion, followed by visuals, speaking, and story. Regarding the rater facet, a wide range of rater severity was identified, but this is what we expected for PA because student raters are different from trained, professional raters. Overall, the students seemed to be able to use the scales well.

About 11% of the raters were reported to have infit values above 1.5, indicating that their ratings were inconsistent. About one third of the raters were reported to have infit values under 0.7, meaning that they used the scales in a limited manner. In fact, these raters used only 4 and 5 on the 5-point Likert scales. Four of them were in the same group, and we found that a majority of presenters in this particular group performed very well and received high scores from the teacher. This means that the four raters, who were reported as the most lenient, did not use the full scales, not because they did not understand them well, but because they probably did not need to. The PA in our project was administered in small groups and is different from PA conducted in a whole class, and this point should be considered when analyzing the results.

The facets analysis reported 39 unexpected responses, but in some cases, raters gave presenters the same scores as the teacher. For example, Rater 311 rated the visuals of Presenter 314 at 1, which was much lower than the expected score of 4.7, but the same as the teacher's. Although we were aware that teachers could be biased to some degree (Lumley & McNamara, 1995), the teacher in our study was considered to be an "ideal" assessor because of experience in teaching presentations and using the rubrics. Thus we treated TA as the criterion with which PA was compared.

Therefore, it could be concluded that Rater 311 was as accurate as the teacher, but because the majority of the raters in the group were not, he was reported as "biased" in the facets analysis. For this reason, we regarded his response as unbiased, and a similar approach was taken to other "unexpected responses" that were actually the same as or closer to the teacher's scores. As a result, only 12 cases remained as biased, 11 of which involved underrating and only one of which involved overrating.

The final step was to examine whether or not these biased responses were related to the four levels of closeness (4 = *very close to him or her*, 3 = *close to him or her*, 2 = *I know him or her a little*, 1 = *I don't know him or her at all*) obtained from the questionnaire (see Appendix A). We conducted a two-way contingency analysis using crosstabs to see if there was a significant interaction between biased scores and the degree of friendship ($p < .05$). We had only 12 biased cases out of 1,026, resulting in no significant results ($\chi^2 = 6.42$, $p = .38$). To sum up, there was no evidence that the degree of friendship influenced PA, meaning that the reliability of PA was not constrained by students' social networks in our study.

Conclusion

In order to develop an effective assessment tool, we conducted this study based on three objectives. First, the hypothesis that personality traits might affect presentation performances was tested. None of the correlations between the Big-Five factors and TA of students' presentations were significant. Second, in order to test the reliability of PA compared with TA, reliability analyses using Cronbach's α and agreement rate and many-facet Rasch analysis were conducted. Overall, intra-class reliability between PA and TA was not high, except for the visuals construct, which showed modest agreement (.74). Further analysis was conducted to check the percentage agreement between PA and TA in high- and low-rated presentations. Though the student raters were more successful at identifying low-rated presentations (66.7%) than high-rated presentations (20%), it was confirmed that there was a still a discrepancy between TA and PA. Third, the facets analysis showed that student raters, overall, assessed peer presentations in a consistent manner; only 12 responses identified as biased cases. The results of the two-way contingency table analysis showed no evidence that friendship between the assessor and the assessed influenced PA.

In conclusion, the students in the present study used the assessment form in a generally unbiased way and made good use of the scales, indicating their understanding of the constructs and providing evidence for their learning. However, in order to achieve better alignment for PA with TA, the form may need to be adapted. A greater amount of training and practice using the form might also bring about greater reliability. In future research, more items from the Big-Five scale of personality traits should be included in the analysis to further test the results of these analyses. Moreover, factors not considered in our research that could have effects on PA could be investigated, such as the effects of peer pressure and students' perceptions of PA. Furthermore, variation in the conditions in which presentations were conducted could also be examined. It is conceivable that students would make different assessments if the presentations were made in front of a larger audience, rather than the relatively small groups in which they were conducted in this study. Finally, it is also plausible that if the presentations were higher stakes assessments in terms of their value in the students' final grades, there might be different results in terms of the effects of levels of closeness between presenters and assessors.

Bio Data

Robert Vaughan is an adjunct lecturer at Rikkyo University. His main research interest is academic second language socialization.

Yukie Saito is an adjunct lecturer at Rikkyo University. Her main research interest is instructed SLA, in particular the role of pronunciation instruction in L2 speech development.

Yukie Saito is a part-time instructor at Waseda University. Her main research areas are high school teachers' cognition, CEFR, and pragmatics.

References

Apple, M. T. (2011). *The big five personality traits and foreign language speaking confidence among Japanese EFL students* (unpublished doctoral dissertation). Temple University, Tokyo, Japan.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York, NY: Routledge.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing, 20,* 89-110.

Cheng, W., & Warren, M. (1997). Having second thoughts: Students perceptions before and after the peer exercise. *Studies in Higher Education, 22,* 233-239.

De Grez, L., Valcke, M., & Roozen, I. (2012). How effective are self- and peer assessment of oral presentation skills compared with teachers' assessment? *Active Learning in Higher Education, 13*(2), 129-142.

Dewaele, J. (2013). The link between foreign language classroom and psychoticism, extroversion, and neuroticism among adult bi- and multilinguals. *The Modern Language Journal, 97,* 670-684.

Field, A. P. (2009). *Discovering statistics using SPSS: And sex and drugs and rock 'n' roll* (3rd ed.). London, UK: Sage.

Garson, G. D. (2008). Path analysis. Statnotes: Topics in multivariate analysis. Retrieved from <http://faculty.chass.ncsu.edu/garson/PA765/path.htm>

Gershon, S. (2008). *Present yourself 2: Viewpoints* (1st ed.). New York, NY: Cambridge University Press.

Harrington, D., & LeBeau, C. (2008). *Speaking of speech: Basic presentation skills for beginners*. Tokyo: Macmillan.

Joyce, M. (2013, May 11). Picking the best intercoder reliability statistic for your digital activism content analysis. *Digital Activism Research Project*. Retrieved from <http://digital-activism.org/2013/05/picking-the-best-intercoder-reliability-statistic-for-your-digital-activism-content-analysis/>

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research, 30,* 411-433.

Larson-Hall, J. (2009). *A guide to doing statistics in second language research using SPSS*. New York; NY: Routledge.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2010, June 1). *Intercoder reliability: Practical resources for assessing and reporting intercoder reliability in content analysis research projects*. Retreived from <http://matthewlombard.com/reliability/>

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12,* 54-71.

Mak, B. (2011). An exploration of speaking-in-class anxiety with Chinese ESL learners. *System, 39,* 202-214.

Namikawa, T., Tani, I., Wakita,T., Kumagai, R., Nakane, A., & Noguchi, H. (2012). Development of a short version of the Japanese big-five scale, and a test of its reliability and validity. *The Japanese Journal of Psychology, 83*(2), 91-99.

Patri, M. (2002). The influence of peer feedback on self- and peer-assessment of oral skills. *Language Testing, 19*(2), 109-131.

Rothstein, M. G., Paunonen, S. V., Rush, J. C., & King, G.A. (1994). Personality and cognitive ability predictors of performance in graduate business school. *Journal of Educational Psychology, 86,* 516-530.

Saito, H. (2008). EFL classroom peer assessment: Training effects on rating and commenting. *Language Testing, 25,* 553-581.

Toyota, H. (2007). *Covariance structure analysis for AMOS.* Tokyo: Tokyo Tosho.

Yamashiro, A. D. (1999). Validating a rating scale using multitrait multimethod analysis. *Temple University Japan Working Papers in Applied Linguistics, Series: Individual Differences in the Japanese EFL Context, 14*. Retrieved from <http://www.tuj.ac.jp/tesol/publications/working-papers/vol-14/yamashiro.html>

Appendix A

Questionnaire to Establish the Degree of Friendship Among Students

このアンケートは最終プレゼンテーションのグループ分けと今後の授業をより良いものにするための研究の一環として使われます。それぞれの学生について該当するものに一つだけチェックをして下さい。このアンケートの結果が成績に影響を与えることは一切ありません。みなさんの情報は保護され、他への利用は一切ありません。ご協力をお願いします。

Student Number	
Grade	
Name	

	Very close to him/her 親しい	Close to him/her まあまあ親しい	I know him/her a little 少し知っている	I don't know him/her at all 全然知らない
Student A				
Student B				
Student C				
Student D				
Student E				

Appendix B

A Short Form of the Japanese Big-Five scale of personality traits

1．全くあてはまらない	Strongly disagree
2．ほとんどあてはまらない	Moderately disagree
3．あまりあてはならない	Disagree a little
4．どちらともいえない	Neither disagree nor agree

5．ややあてはまる　Agree a little
6．かなりあてはまる　Moderately agree
7．非常にあてはまる　Strongly agree

Item	No.	Japanese	English translation	Agree/Disagree
Neuroticism	N1	悩みがち	worrying	1 2 3 4 5 6 7
	N2	不安になりやすい	unstable	1 2 3 4 5 6 7
	N3	心配性	anxious	1 2 3 4 5 6 7
	N4	気苦労の多い	having cares	1 2 3 4 5 6 7
Extraversion	S1	話し好き	talkative	1 2 3 4 5 6 7
	S2	陽気な	cheerful	1 2 3 4 5 6 7
	S3	外向的	outgoing	1 2 3 4 5 6 7
	S4	社交的	sociable	1 2 3 4 5 6 7
Openness	O1	独創的な	original	1 2 3 4 5 6 7
	O2	多才の	versatile	1 2 3 4 5 6 7
	O3	進歩的	progressive	1 2 3 4 5 6 7
	O4	洞察力のある	insightful	1 2 3 4 5 6 7
Agreeableness	A1	温和な	gentle	1 2 3 4 5 6 7
	A2	寛大な	generous	1 2 3 4 5 6 7
	A3	親切な	kind	1 2 3 4 5 6 7
	A4	良心的な	good-natured	1 2 3 4 5 6 7
Conscientiousness	C1	計画性のある	planned	1 2 3 4 5 6 7
	C2	勤勉な	hardworking	1 2 3 4 5 6 7
	C3	几帳面な	organized	1 2 3 4 5 6 7
	C4*	いい加減な	irresponsible	1 2 3 4 5 6 7

Note. 20 Japanese items were taken from Namikawa et al. (2012) and translated into English for reference by the researchers. N1, S1, A1, and C1 mean the first item from Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness, respectively. C4 was a reverse item because this item carries a negative meaning and others have positive meanings.

Appendix C
Assessment Form

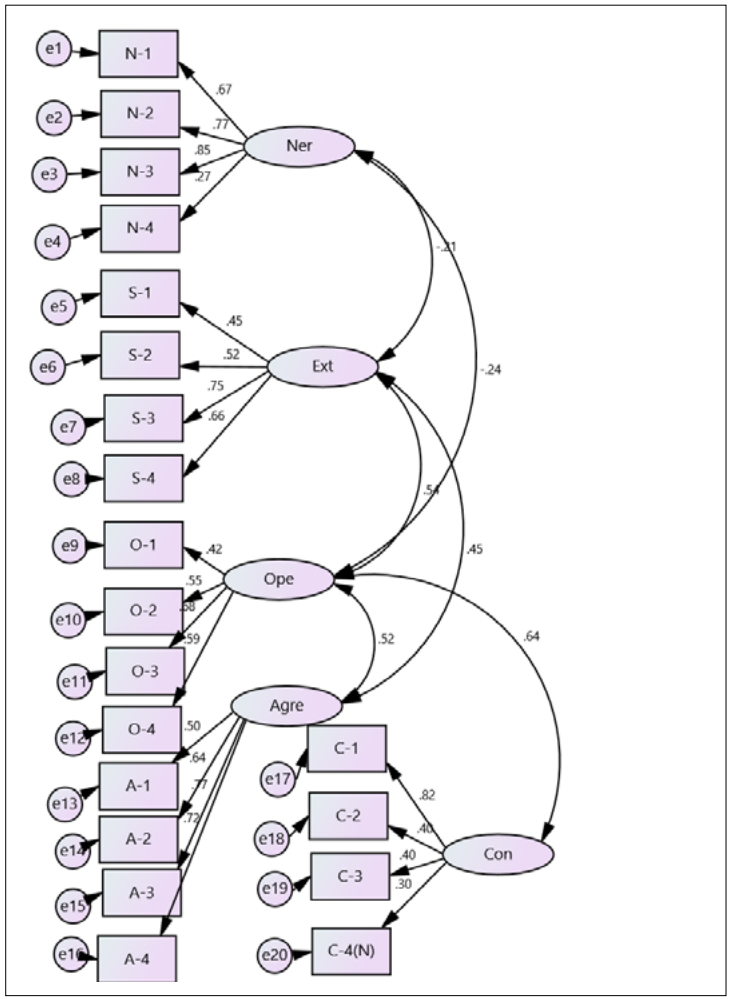
Please evaluate your group's presenters using the information below. Please also write a self-evaluation.

Time				
4-minutes +	3.5-minutes +	3-minutes +	2-minutes +	2-minutes -
5	**4**	**3**	**2**	**1**
Memory / looking at notes				
The whole speech from memory.	Very rarely checking notes.	Mostly from memory, but sometimes clearly reading.	Mostly reading, but sometimes looking up.	Only reading.
5	4	3	2	**1**
Speaking / physical performance				
Excellent voice quality, volume, eye contact, and gestures.	Good voice quality, volume, eye contact and gestures.	Mostly good voice quality, volume, eye contact, and gestures.	Often not good voice quality, low volume, little eye contact, and few gestures.	Poor voice quality, cannot hear, no eye contact, and no gestures.
5	**4**	**3**	**2**	**1**
Story / content				
Very interesting, easy to understand everything, and very good introduction/ body/conclusion.	Interesting, almost everything was easy to understand, and good introduction/body/ conclusion.	Mostly interesting, mostly easy to understand, and satisfactory to good introduction/body/ conclusion.	Not very interesting at times, sometimes difficult to understand, and short introduction/ body/ conclusion.	Boring at times, often difficult to understand, and no introduction /body/ conclusion
5	4	**3**	**2**	**1**

Visuals				
Very carefully prepared, simple/easy to read/see, attractive, relevant to/ supporting the presentation, and used well.	Carefully prepared, simple/ easy to read/ see, moderately attractive, mainly relevant to / supporting the presentation, and usually used well.	Some effort taken to prepare, mainly simple / easy to read/see, okay to look at, usually relevant to/ supporting the presentation, and usually used moderately well.	Little effort taken to prepare, sometimes complicated / difficult to read/see, not very attractive, some parts not relevant to/not supporting the presentation, and not always used well.	Very little effort made to prepare, too complicated, very difficult to read/see, unattractive, not relevant / does not support the presentation, or not used / used badly.
5	4	3	2	1

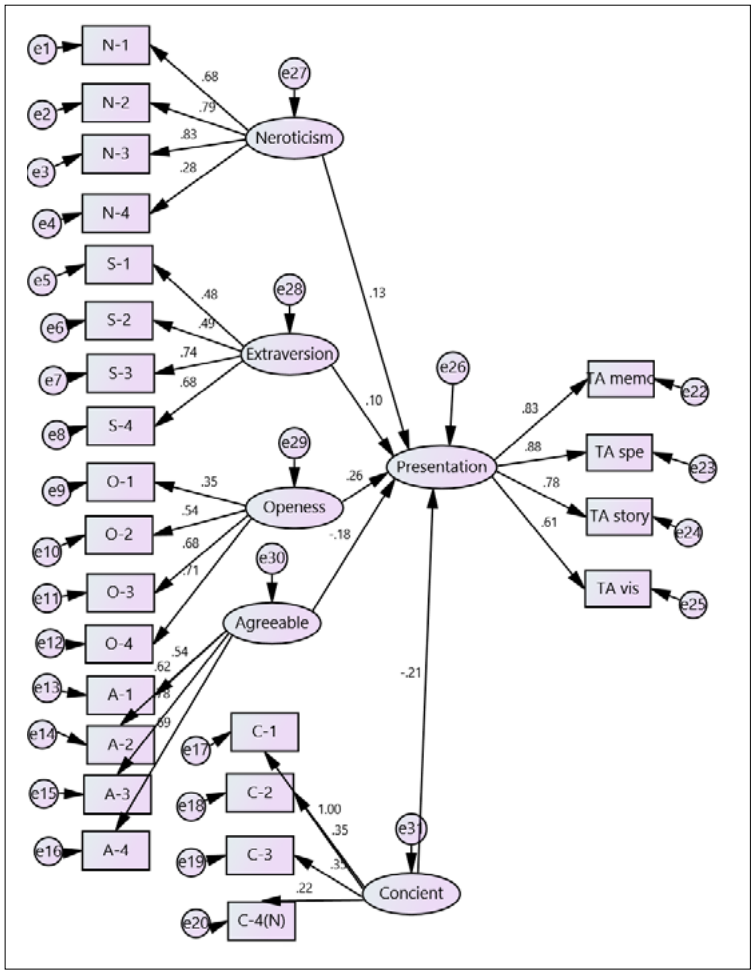
No.	Name	Time	Memory / looking at notes	Speaking / physical performance	Story / Content	Visuals	Total	Teacher total	Average
301									
302									
303									
304									
305									
306									
307									
308									
309									
310									

Appendix D
The Result of the Confirmatory Factor Analysis



Note. Ner = neuroticism; Ext = extraversion; Ope = openness; Agre = agreeableness; Con = conscientiousness. N1, S1, A1, and C1 = the first item from neuroticism, extraversion, openness, agreeableness, and conscientiousness, respectively.

Appendix E
The Result of the Covariance Structure Analysis



Note. TA memo = teacher's assessment of memory / looking at notes; TA spe = teacher's assessment of speaking / physical performance, TA story = teacher's assessment of story / content, TA vis = teacher's assessment of visuals; Neroticism = neuroticism; Openess = Openness; Agreeable = agreeableness; Conscient = conscientiousness.