2020年（令和2年度）博士学位論文

学生による相互評価力とプレゼンテーション力に及ぼす要因に関する実証的研究
―学生の相互評価と教員による評価との相関と評価者トレーニングに基づいて―

京都外国語大学大学院外国語学研究科　博士後期課程

異言語・文化専攻

言語教育領域

2015DC0001

笠巻知子

目次

Abstract

In recent years, there has been a tendency to emphasize communication in English education. Students actively communicate in English in a class. However, the evaluation is only an evaluation for the grade. As it is time-consuming for instructors to assess students' speaking abilities. (Luoma, 2004; Ur, 2012), it is not re-evaluated after a day to ensure the evaluator's reliability. Therefore, it is hard to say that one evaluator's evaluation is reliable (Hughes, 1989).

One of the solutions to these problems for evaluation is the application of peer assessment. It tends to be adopted more and more (Goh & Burns, 2012; Fukazawa, 2010; Fujiwara et al., 2007b; Falchikov & Goldfinch, 2000; Luoma, 2004) to and to supplement the teacher assessment and to clarify the learning objectives to the learners. If peer assessment can be utilized, it will be possible to evaluate students speaking more objectively from various perspectives. Peer assessment can be conducted during class hours, which may reduce the time and effort required for teacher assessment (Fukazawa, 2009; Fukazawa, 2010: Okuda & Otsu). 2010).

However, to incorporate peer assessment, it is necessary to verify its reliability. There are conflicting views between the theory that peer assessment is reliable (Fukazawa, 2010; Nakamura, 2002; Okuda & Otsu, 2010) and the theory that it is unreliable (Freeman, 1995; Kasamaki, 2016).

Since 2016, the author has examined the reliability of peer assessment by Japanese university students, what affects their evaluation ability, and whether peer assessment can supplement the teacher assessment.

This study aimed to examine the reliability of peer assessments by Japanese University students in order to supplement to the evaluation by teachers. It was considered that the student's "presentation skills," "English proficiency," "prior knowledge," and "rater training" influence their evaluation skills as raters. The author examined the correlation between the teacher assessment and peer assessment. In addition, some tendencies for

peer assessment were examined.

The main purpose of this study is to report the empirical studies investigating what affect students' ability to rate presentations delivered by peers, whether rater training can improve their evaluation ability, and whether rater training can be one of the guidance for oral presentation.

In this study, six research questions (RQ) were set up and verified.

RQ 1 : Do students' oral presentation skills affect their evaluation ability?

RQ 2 : Do students' English proficiency level affect their evaluation ability?

RQ 3 : Do students' prior knowledge affect their evaluation ability?

RQ 4 : Does rater training affect students' evaluation ability?

RQ 5 : Does rater training by each evaluation item affect students' evaluation ability?

RQ 6 : Does rater training affect students' oral presentation skills?

This thesis consists of nine chapters.

Ch.1 Introduction

Ch.2 Previous research

Ch.3 Do students' oral presentation skills affect their evaluation ability?

Ch.4 Do students' English proficiency level affect their evaluation ability?

Ch.5 Do students' prior knowledge affect their evaluation ability?

Ch.6 Does rater training affect students' evaluation ability?

Ch.7 Does rater training by each evaluation item affect students' evaluation ability?

Ch.8 Does rater training affect students' oral presentation skills?

Ch.9 Conclusion

Chapter 1 explains the background of this study, the reasons why the author attempted this study, and the maim goal. Chapter 2 surveys the previous research about speaking assessment, the essential concepts of peer assessment, the advantages and disadvantages of peer assessment, the

reliability of peer assessment, and empirical studies examining the reliability of peer assessment. Chapter 2 also reports the summary of Kasamaki (2016) and what studies should be examined and explored. Chapter 3 reports whether students' oral presentation skills affect their evaluation ability as raters. In addition, some tendencies for peer assessment were examined. Chapter 4 reports whether students' English skills affect their evaluation ability as raters. In addition, some tendencies for peer assessment were examined. Chapter 5 reports whether students' prior knowledge about the topics of presentations delivered by peers affect their evaluation ability as raters. In addition, some tendencies for peer assessment were examined. Chapter 6 reports whether rater training affects their evaluation ability as raters. Chapter 7 reports whether rater training for each evaluation item affects their evaluation ability as raters. Chapter 8 reports whether rater training affects students' oral presentation skills. Lastly, chapter 9 summarizes the results of the six studies from chapter 3 to 8 and further studies.

The details of each chapter are going to be treated as follows.

Chapter 1 described the evaluation methods used in the recent speaking classes at Japanese universities, their problems, and also explained the background and motives for conducting this research.

Chapter 2 introduced previous research on speaking assessment, especially, the difficulty of speaking assessment, evaluation methods, and factors that influence the evaluation. Also, the author has summarized previous research on the advantages and disadvantages of peer assessment and empirical research that verified the reliability of peer assessment.

The advantages of peer assessment are that it clarifies the learning goals (Goh & Burns, 2012; Fukazawa, 2010; Luoma, 2004), and increases learners' motivation for learning (Nakamura, 2002). Students can learn by evaluating presentations each other (Luoma, 2004; Fujiwara et al., 2007b).

On the other hand, the disadvantages are that students tend to be more

lenient compared to teachers and that the range of evaluation is narrow, avoiding extreme evaluation (Cheng & Warren, 2005; Hughes & Large, 1993; Fukazawa, 2010; Kasamaki, 2016; Otoshi & Hefferman, 2007).

In studies that verified the reliability of peer evaluation, there is a conflict between the theory that student evaluation is reliable and incorporated into a part of the grades (Fukazawa, 2010 ; Luoma, 2004 ; Nakamura, 2002 ; Okuda $ Otsu, 2010) and the theory that student evaluation is not reliable (Freeman, 1995 ; Hirai *et al*, 2011 ; kasamaki, 2016 ; Oi, 2012). Besides, the author summarized what has been clarified in the past research and the research issues to be conducted.

Chapter 3 reports whether students' oral presentation skills affect their evaluation skills as raters. 61 university students participated in this study. Participants assessed presentations delivered by their peers. The author examined the correlation between peer assessment and teacher assessment with reference to the participants' presentation skills. The teacher evaluated each student simultaneously with the students. Based on the teacher evaluation, the students were divided into upper, middle, and lower groups. The correlation between teacher assessment and peer assessment of each grade group (upper, middle, lower) was examined for each performance. A moderate overall correlation was found between peer assessment and teacher assessment ($r = .529 \sim .606$).

Then, the author investigated whether there was a difference in the students' evaluation ability of each grade group. Differences were observed between the upper and middle groups and between the upper and lower groups. The effect size was small. Since there was no difference between the middle group and the lower group, the evaluation ability of the upper group was the lowest. These results indicate that the upper group students, who are said to have the highest presentation ability, have the lowest evaluation ability, and that the students' evaluation ability as raters was not related to their presentation skills.

The author also examined the correlation between peer assessment and teacher assessment with reference to each evaluation items. A weak correlation was found especially in the items of "research" and "originality" related to the content of the presentation.

In addition, some tendencies for peer assessment were examined. In general, students tended to be more lenient compared to teacher assessment. The upper group students tended to give a lower evaluation, while the lower group students tended to give a higher evaluation. They also tend to evaluate the presentation by peers based on their own performance.

Chapter 4 reports whether the students' English proficiency affects their evaluation ability as an evaluator. Based on the students' TOEIC IP scores, the students were divided into upper, middle, and lower English proficiency groups. The correlation between teacher assessment and peer assessment of each grade group (upper, middle, lower) was examined for each performance. A moderate overall correlation was found between peer assessment and teacher assessment (r = .551~589) with little difference among the three group. It was thought that higher ability group more closely to the teacher, but it was confirmed that the middle English proficiency group graded most closely to the teacher. These results indicate that the rating ability of students was not related to their English proficiency. In addition, some tendencies for peer assessment were examined. In general, students tended to be more lenient compared to the teacher. Also, students tended to change their evaluation depending on their relationship with their classmates because they feel awkward to evaluate to classmates. It was considered that personality influenced the evaluation.

Chapter 5 reports whether students' prior knowledge affects the student's evaluation ability as an evaluator. Based on the degree of students' prior knowledge, the students were divided into 5 groups : background 2, background 4, background 6, background 8, and background 10. Some tendencies for each grade of their prior knowledge were examined. These

results indicate that some of the student's prior knowledge may have a slight effect on the student's evaluation tendency. In general, students tended to be more lenient compared to teacher. Also, students to students tend to give evaluations that were close to their own perceived level of performance. Psychological factors such as the inexperience of the student's evaluation, and the awkwardness of evaluating to classmates was considered.

To utilize peer assessment, it is necessary to improve students' evaluation ability and increase the reliability of peer evaluation. For that purpose, evaluation items, judgment criteria, and viewpoints of evaluation should be clearly indicated on the evaluation sheets. Also, it was considered necessary to conduct rater training using rubrics.

Chapter 6 reports whether rater training affects students' evaluation ability as evaluators. Participants in this study included 61 university students. They were divided into two groups. One group (hereinafter *Shosai* group) was given full-scale rater training twice. The other group (hereinafter *Kani* group) was given simple-procedure rater training twice.

The first rater training was conducted during class one week before the midterm presentation. The second one was conducted during class one week before the final presentation. For the training, the author prepared rubrics, audio demos of "pronunciation" and "pause," slide samples, and student presentation videos. The rubrics was created by the author based on the evaluation sheet used in the "Project-based English program." The time required for rater training for *Shosai* group was about 45 minutes , whereas for *Kani* group about 15 minutes.

After completing the first rater training, the correlation between teacher assessment and peer assessment of each group (*Shosai, Kani*) was examined. A weak overall correlation of both groups was found between peer assessment and teacher assessment ($r = .383$).

This result indicates that only once rater training does not affect students' evaluation ability. The difference in the quality and content of

one-time rater training does not affect students' evaluation ability.

The author then, investigated the influence of the number of rater training on students' evaluation ability. After the second rater training, the correlation between teacher assessment and peer assessment of each group was examined. A moderate overall correlation of both *Shosai* group ($r = .619$) and *Kani* group ($r = .625$) was obtained. This result indicates that the total evaluation ability of both groups was improved after the second rater training.

Then, the author investigated whether students' evaluation ability in each group was improved. The students' evaluation ability after the first evaluation training and the second evaluation training was examined. In the "total" of the *shosai* group evaluation, there was a difference between the evaluation ability after the first training and the evaluation ability after the second training. The effect size was large. The students' evaluation ability after the second time rater training was significantly higher than that after the first training. In other words, the evaluation ability of students can be improved by conducting full-scale evaluator training twice. Also, in the "total" of the evaluation of the *Kani* group, there was a difference between the evaluation ability after the first training and the evaluation ability after the second training. The effect size was large. Even a simple-procedure rater training could improve the students' evaluation ability by conducting it twice.

However, there was no difference in the two groups' overall evaluation ability. Therefore, the number of implementations rather than the content and quality of the rater training may have affected the student's evaluation ability.

The ability to evaluate "eye contact," "pronunciation," and "pose." can be improved by conducting full-scale rater training twice. However, the ability to evaluate "contents," can be improved by conducting simple-procedure training twice. The evaluation ability for "content" is related to other factors

such as prior knowledge of the content of the presentation and English ability to fully understand the content, rather than the quality of the rater training.

Conducting the simple-procedure rater training twice was able to improve *Kani* group's overall evaluation ability. On the other hand, it was not possible to improve the evaluation ability for all evaluation items except "contents."

These results indicate that the lack of students' evaluation ability due to inexperience may have been improved by increasing the number of times of training. By repeating the training more than once, the evaluation ability of the students is likely to be improved. Therefore, rater training may not only improve students' evaluation ability but also increase the reliability of peer evaluation.

Since it takes a certain amount of time to carry out rater training during class hours, there remains the problem that the method should be reviewed to enable more feasible rater training in the classroom.

Chapter 7 investigated whether rater training for each evaluation item affects students' evaluation ability as evaluators. Participants in this study included 79 university students. They were divided into two groups. One group(hereinafter *Shosai* group) was given rater training for each evaluation item for three weeks before the peer evaluation. The other group (hereinafter *Kani* group) was given the full-scale rater training described in Chapter 6.

The author narrowed down the evaluation items to "presentation" and further subdivided them into "posture", "eye contact", "gesture", "volume", "clarity", and "slide". Based on this, the rubric has been changed. The procedure of rater training was same as conducted in Chapter 6.

The time required for the rater training for *Shosai* group was about 30 minutes in the detailed group, whereas for *Kani* group training about 45 minutes.

As a result, a moderate correlation was obtained for *Shosai* group (r

= .699) and the *kani* group (r = .698) in the total evaluation score. The overall evaluation ability of the students was not high in either the *Shosai* group or the *Kani* group. As for each evaluation item, the correlation varied. Therefore, rater training for evaluation item does not affect students' evaluation ability.

It took much time to train all the evaluation items for the students once just before the evaluation. On the other hand, rater training for each evaluation item could reduce the time required for one training. It seems that rater training for each evaluation item could compensated the lack of students' evaluation ability due to the inexperience of evaluation to some extent, but as a result, it did not lead to the improvement of the evaluation ability for all evaluation items. The need to review the implementation method remained as a future issue.

Chapter 8 reports whether rater training not only improves the student's evaluation ability as an evaluator but also whether rater training can improve students' oral presentation skills. The students who did not receive rater training were designated as the "group without rater training,(hereinafter control group)" and the students who received rater training were designated as the "group with rater training (hereinafter experimental group)." To investigate whether rater training affects students' oral presentation skills, the author compared the results of the midterm and final presentations in each group. As a result, there was no difference in the growth of grades from the midterm presentation to the final presentation in both groups. Rater training is unlikely to affect students' oral presentation skills.

In addition, in control group, there was no difference in the grades from the midterm presentation to the final presentation On the other hand, in experimental group, the grades of the final presentation were slightly higher than that of final presentation. This result indicates that rater training may have a slight effect on students' oral presentation skills.

Also, as the result of investigating whether the difference in the content of the rater training affects the student's oral presentation skills, the simple-procedure rater training might affect the student's oral presentation skills.

Based on these results, in order for the rater training to affect students' oral presentation skills, the evaluation items, the method of rater training, the number of times of implementation, and the timing of implementation should be reviewed, and training with simplified procedures should be continued.

In chapter 9 summarizes the results and conclusions of six studies in this thesis.

In this study, the author has investigated the factors that affect student evaluation ability to examine the reliability of peer evaluation.

As a result, it was found that the students' "presentation skills," "English proficiency," and "prior knowledge" did not affect the students' evaluation ability. Then, when the evaluator training was conducted, the students' lack of evaluation ability could be compensated to some extent by the training and also, evaluator training could have a slight impact on students' presentation skills.

Also, rater training is not merely to increase the degree of agreement with teacher evaluation. Through rater training, students will be able to understand the quality of the presentation. Furthermore, they will learn good presentations and bad presentations delivered by their classmates by peer evaluation. This leads to become autonomous learners in the future.

Finally, conducting rater training and learning about evaluation items that are difficult for students to evaluate can be a good opportunity to review the teachers' evaluation. In other words, it can be said that it improves the evaluation ability of students and improves the evaluation ability of teachers.

As further research, it was impossible to clearly identify the factors that

influence students' evaluation ability as evaluators in this research, so it is necessary to continue the research to investigate the factors. In addition, even if the rater training was conducted twice, the students' evaluation ability for the assessment criteria related to the presentation's content was not improved. Therefore, what will affect these evaluation items in the future is needed to be explored. Then, it is necessary to review the improvement points of the rater training clarified in this study and verify whether the students' evaluation ability of the students can be improved by continuously implementing the training with simplified procedures. Finally, it is necessary to verify whether the rater training with speech training can improve the students' oral presentation skills.

抄録

近年の英語教育においては，コミュニケーション重視の傾向にあり，教室内においても学生が積極的にコミュニケーションする姿が見受けられる。しかし，学生のスピーキング・パフォーマンスを評価するのは，担当教員が成績のために一人で行うのが現状である。学生一人ひとりの評価を担当教員一人で行うのには，相当な時間と労力が必要とされることから（Luoma, 2004; Ur, 2012），日を置いて再度評価し，評価者信頼性を確保していないことが多く，スピーキングの評価に関してはあまり適切に行われていない。そのため，一人の評価者による評価は，信頼性があるとは言い難い（Hughes, 1989）。

評価に対するこれらの問題点の解決策の一つとして，学生による相互評価の活用があげられる。教員による評価を補足するために，また学習者に学習目標を明確にするために，スピーキングのクラスにおいて，ますます取り入れられる傾向にある（Goh & Burns, 2012; Fukazawa, 2010; 藤原他,2007b; Falchikov & Goldfinch, 2000; Luoma, 2004）。また，学生による相互評価は授業時間内に行うことができるため，教員の評価にかかる時間や労力が少しでも軽減されるとも考えられる（深澤, 2009；Fukazawa, 2010：Okuda ＆ Otsu, 2010）。

しかし，学生による相互評価を取り入れるには，その信頼性を検証する必要があり，これまで多くの教員や研究者により様々な研究が行われてきた。その結果，学生による相互評価は信頼できるとする説（Fukazawa, 2010; Nakamura, 2002; Okuda & Otsu, 2010）と，信頼できないとする説（Freeman, 1995; 笠巻, 2016）と，相反する見解が見受けられた。しかし，信頼できないとする場合，何が原因なのかを検証した報告は見当たらなかった。

筆者は，日本人大学生を対象に，何が学生の評価力に影響を及ぼすのか，また，学生による相互評価は教員による評価を補足するものとなり得るのか，その信頼性を2016年以来検証してきた。本論文では，学生による相互評価の信頼性を検証し，スピーキングの内容を評価するにあたり，学生の視点を取り入れられるかどうかを検討した。評価者としての学生の評価力には，学生の「プレゼンテーション力」，「英語力」，「予備知識の有無」，「評価者トレーニング」が影響を及ぼす要因であることが考えられた。本論文では，これらの4つの要因が評価者としての学生の評価力に影響を及ぼすかを調べ，学生による相互評価の信頼性を検証することを試みた。

さらに，評価者トレーニングを，プレゼンテーションの指導の一環として評価項目別にトレーニングをすることで，学生の評価力に影響を及ぼすかについて調べた。また評価者トレーニングが学生の評価力だけではなく，学生のプレゼンテーション力に影響を及ぼすかについて調べ，プレゼンテーションの指導の一つとなり得るかについ

て検証を行った。

本論文は，次の 9 章から構成されている。

第 1 章　はじめに
第 2 章　先行研究
第 3 章　研究 1 : 学生の「プレゼンテーション力」が評価者としての学生の評価力に影響を及ぼすか？
第 4 章　研究 2 : 学生の「英語力」が評価者としての学生の評価力に影響を及ぼすか？
第 5 章　研究 3 : 学生の「予備知識」が評価者としての学生の評価力に影響を及ぼすか？
第 6 章　研究 4 :「評価者トレーニング」が評価者としての学生の評価力に影響を及ぼすか？
第 7 章　研究 5 :「評価項目別評価者トレーニング」が評価者としての学生の評価力に影響を及ぼすか？
第 8 章　研究 6 :「評価者トレーニング」が学生の「プレゼンテーション力」に影響を及ぼすか？
第 9 章　終わりに

第 1 章では，最近の日本の大学でのスピーキング・クラスにおいて行われている評価方法，また指摘されている問題点について述べ，本研究を行った背景および動機について述べる。

第 2 章では，スピーキングの評価の観点，評価方法，評価に影響を及ぼす要因についての先行研究を紹介し，学生による相互評価の利点および欠点についての先行研究と，相互評価の信頼性を検証した実証的研究をまとめる。

第 3 章から第 6 章では，評価者としての学生の評価力に影響を及ぼすと考えられる 4 つの要因，学生の「プレゼンテーション力」,「英語力」,「予備知識」そして「評価者トレーニングの有無」が評価者としての学生の評価力に影響を及ぼすかについて，教員による評価と学生による相互評価の相関を調べ，検証する。また，各要因別に見た，学生の評価傾向についても報告する。

第 7 章では，評価者トレーニングを，到達目標となる評価項目別に行い，実際の相互評価の直前だけではなく，授業の一部に組み入れることで，学生の評価力に影響を及ぼすかについて検証する。

第 8 章では，評価者トレーニングを受けた学生と，受けなかった学生それぞれのプレゼンテーションの成績の伸びを比較し，評価者トレーニングがプレゼンテーション

力向上のための指導の一つとなり得るか，その有効性について報告する。さらに，第7 章で行った評価者トレーニングの内容の違いが，学生のプレゼンテーション力に影響を及ぼすかについても報告する。

第 9 章では，第 3 章から第 8 章までの 6 つの研究から得られた結果をまとめ，それらの研究の問題点と今後の研究の課題について述べる。

以下に各章の詳細について述べる。

第 1 章では，近年の大学におけるスピーキングの授業では，学生のスピーキング力がどのように評価されているのか，その現状と指摘されている問題点について説明した。また，その問題点の解決策の一つとして考えられる学生による相互評価は，信頼できるとする説と信頼できないとする説があるため, 学生による評価を活用するには, その信頼性を検証する必要がある。しかし，信頼できないとする場合，何が原因なのかを検証した報告は少ない。評価者として学生の評価力にはプレゼンテーションの内容を十分に理解できるだけの英語力，その内容に対する予備知識の有無，プレゼンテーション力，そして評価者トレーニングの 4 つの要因があると考えられたことから，本論文では，下記の 6 つのリサーチ・クエスチョン（RQ）を立て，検証を行った。

RQ.1 学生の「プレゼンテーション力」が評価者としての学生の評価力に影響を及ぼすか

RQ.2 学生の「英語力」が評価者としての学生の評価力に影響を及ぼすか

RQ.3 学生の「予備知識」が評価者としての学生の評価力に影響を及ぼすか

RQ.4「評価者トレーニング」が評価者としての学生の評価力に影響を及ぼすか

RQ.5「評価項目別評価者トレーニング」が評価者としての学生の評価力に影響を及ぼすか

RQ.6 評価者トレーニングは学生のプレゼンテーション力に影響を及ぼすか

第 2 章では，スピーキングの評価の難しさ，スピーキングの評価方法，そして，評価に影響を及ぼす要因についてまとめた。さらに，最近多くの授業で取り入れられる傾向にある，学生による相互評価について，その利点，欠点，そして信頼性を検証した先行研究についてまとめた。相互評価の利点としては，学習者に学習目標を明確にさせる（Goh & Burns, 2012; Fukazawa, 2010; Luoma, 2004），学習者の学習に対する動機づけが高まる（Nakamura, 2002），そして，相互評価を通して，学生がお互いに学び合うことができる（Luoma, 2004; 藤原他, 2007b）といった点などが挙げられる。一方，欠点としては，教員による評価よりも若干甘くなる傾向がみられることと，極端な評価を避け，評価の幅が狭いことなどが挙げられる (Cheng & Warren, 2005; Hughes & Large, 1993; Fukazawa, 2010; 笠巻, 2016; Otoshi & Hefferman, 2007)。相互評価の信頼性

を検証した研究では，学生による評価は信頼でき，成績の一部に組み入れられるとする説と，教員による評価との一致度が低く，信頼性に欠けているとする相反する意見が見られた。

第 3 章では，評価者としての学生の評価力に影響を与えるとされる 4 つの要因のうち，学生の「プレゼンテーション力」が学生の評価者としての評価力に影響を及ぼすかを調べた。各学生のプレゼンテーション力を偏差値化し，偏差値 55 以上を上位群，偏差値 45 から 54 を中位群，偏差値 45 未満を下位群とし，学生が行ったプレゼンテーションに対する学生による相互評価と，教員による相互評価との相関係数を成績群ごとに算出した。その結果，評価の合計点において，上位群，中位群，下位群すべてにおいて，中程度の相関(r = .529〜.609)が得られた。どの成績群の評価力も高くないことが確認された。そして，学生の評価力が，プレゼンテーション力の成績群間に差があるかを調べた結果，上位群と中位群，また上位群と下位群との間には差が見られ，効果量はそれぞれ小であり，中位群と下位群との間には差が見られなかったことから，上位群の評価力が一番低いことがわかった。プレゼンテーション力が一番と高いとされる上位群の学生の評価力が一番低いことが確認されたことから，学生のプレゼンテーション力は，評価者としての学生の評価力には影響を及ぼさない可能性が高いことがわかった。

評価項目別に見ると，どの群においても，特にプレゼンテーションの内容にかかわる「リサーチ」と「オリジナリティ」の項目における，教員による評価と学生による相互評価との相関係数が低くなることがわかった。また，学生の評価傾向として，学生全体としては，教員の評価と比べると甘めの評価を行うことがわかったが，成績群別にみると，上位群の学生は低めの評価を行う傾向があり，逆に下位群の学生は高めの評価を行う傾向があることがわかった。その要因として，学生は評価を行う際に，自分自身のプレゼンテーションの出来と比較していることが考えられた。

第 4 章では，評価者としての学生の評価力に影響を及ぼすと考えられる 4 つの要因の一つである，学生の「英語力」が評価者としての学生の評価力に影響を及ぼすかを調べた。各学生の英語力を、TOEIC IP スコアに基づいて偏差値化し、偏差値 55 以上を上位群，偏差値 45 から 55 未満を中位群，偏差値 45 未満を下位群とし，学生が行ったプレゼンテーションに対する学生による相互評価と，教員による相互評価との相関係数を成績群ごとに算出した。その結果，評価の合計点において，上位群，中位群，下位群すべてにおいて，中程度の相関 (r = .551〜.589)が得られたことから，どの成績群の評価力も高くないことが確認された。そして，学生の評価力が，英語力の成績群間に差があるかを調べた結果，英語力が高いとされる上位群ではなく，中位群の評価力が一番高いことが確認されたことから，学生の英語力は，評価者としての学生の評価力には影響を及ぼさない可能性が高いことがわかった。第 3 章の結果と同じように，

どの群においても，特にプレゼンテーションの内容にかかわる「リサーチ」と「オリジナリティ」の項目における，教員による評価と学生による相互評価との相関係数が低くなることが第 4 章でも確認された。また，学生の評価傾向として，学生全体として，教員による評価と比べると甘目の評価を行う傾向があることがわかった。その要因として，クラスメイトに対して評価することに気まずさを感じることから，高めの点数をつけてしまうといった心理的な要因や，他者との関係性で評価を変えてしまうといった評価者の性格が，評価に影響を及ぼしていると考えられた。

第 5 章では，評価者としての学生の評価力に影響を及ぼすと考えられる 4 つの要因の一つである,「予備知識」が評価者としての学生の評価力に影響を及ぼすかを調べた。学生の予備知識の高低の違いによって，学生の評価傾向は変わるかを調べるために，クラスメイトの発表内容についてどのくらい知っていたかについて，学生の予備知識の程度を 2 から 10 の 5 段階に分け, 予備知識の程度別に学生一人一人の発表に対する，教員による評価の点数と学生による相互評価の点数の差を調べた。その結果，学生の予備知識の多少は，学生の評価傾向に少しの影響を及ぼす可能性があることがわかった。また，学生の評価傾向として，教員による評価と比べると，学生による評価は甘目になることがわかった。その要因としては，クラスメイトを評価することに対する心理的な要因，学生の評価経験の浅さが関係していることが考えられた。学生の評価傾向として，これまでの先行研究で報告されているように，教員による評価と比べて甘目の評価を行うことが本研究でも確認された。その理由として，学生の評価経験の浅さ, それ故に, 評価の際に自分自身のプレゼンテーションの出来と比較してしまう，さらに，クラスメイトに対して評価することへの気まずさといった心理的な要因が考えられた。

学生による相互評価を活用するためには，学生に評価力をつけ，評価の信頼性を高めることが必要であり，そのためには，評価項目，判定基準および評価の観点が明確に示されたルーブリックを用いて，評価者トレーニングを行うことが必要であると考えられた。

第 6 章では，評価者としての学生の評価力に影響を及ぼすと考えられる 4 つの要因の一つである，評価者トレーニングが評価者としての学生の評価力に影響を及ぼすかを調べた。本格的な評価者トレーニングを行う学生(29 名)を詳細グループと，簡易的な評価者トレーニングを行う学生（32 名）を簡易グループとし，中間発表の 1 週間前の授業時に 1 回目，最終発表の 1 週間前の授業時に 2 回目の評価者トレーニングを行った。トレーニングには,「プロジェクト発信型プログラム」で使用されている評価シートに基づき筆者が作成したルーブリック,「発音」「ポーズ」の音声デモ，スライドのサンプル，学生の発表ビデオを使用した。評価者トレーニングに要した時間は，詳細グループで約 45 分間，簡易グループで約 15 分間であった。

評価者トレーニングの 1 回目を行った後の，教員による相互評価と学生による相互評価の相関係数を，詳細グループ，簡易グループごとに算出した結果，評価の合計点において，両グループともに，弱い相関が得られた（r = .383）。つまり，評価者トレーニングの 1 回目後の総合的な評価力は，詳細グループ，簡易グループのどちらも高くないことが確認された。評価力を高めるための評価者トレーニングが，1 度の評価者トレーニングの質や内容の違いでは，学生の評価力に影響を及ぼすとは言えないことがわかった。

次に，評価者トレーニングの回数が学生の評価力に影響を及ぼすかを調べるために，評価者トレーニングの 2 回目を行った後の，教員による相互評価と学生による相互評価の相関係数を，詳細グループ，簡易グループごとに算出した。その結果，評価の合計点において，詳細グループにおいては，中程度の相関が得られ（r = .619），簡易グループにおいても，評価の合計点において，中程度の相関が得られた（r =.625）。このことから，評価者トレーニングの 2 回目を行った後の合計的な評価力は，詳細グループ，簡易グループのどちらも高くなったことが確認された。

そして，各グループにおける評価力の伸びを調べるために，評価者トレーニングの 1 回目を行った後と，2 回目を行った後の学生の評価力の差を調べた結果，詳細グループの評価の「合計」においては，トレーニングの 1 回目後の評価力と 2 回目後の評価力の間に差があり，効果量は大で，1 回目後より 2 回目後の学生の評価力は有意に伸びていることがわかる。つまり，本格的な評価者トレーニングを 2 回行うことで学生の評価力を高めることができると言えよう。

簡易グループの評価の「合計」においては，トレーニングの 1 回目後の評価力と 2 回目後の評価力の間に差があり，効果量は大で，簡易グループにおいても，1 回目後より 2 回目後の学生の評価力は有意に伸びていることがわかる。つまり，簡易なトレーニングであっても，2 回実施することで，学生の評価力を高められる可能性があることがわかった。本格的な評価者トレーニング，簡易なトレーニングをそれぞれ 2 回行っても，両グループの全体的な評価力に差はでなかったことから，評価者トレーニングの内容や質よりも，2 回というトレーニングの実施回数の方が学生の評価力に影響を及ぼした可能性がある。

しかし，評価項目別に見ると，「アイコンタクト」，「発音」，「ポーズ」においては，本格的な評価者トレーニングを 2 回行うことで，これら 3 つの項目に対する評価力を上げることができると言えよう。しかし，「内容」においては，本格的な評価者トレーニングを 2 回行った詳細グループより，簡易的なトレーニングのみ 2 回行った簡易グループの評価力の方が上回る結果となった。これは，「内容」の評価に対する評価力は，評価者トレーニングの質よりも，発表内容に対する予備知識や，内容を十分理解できるだけの英語力といった他の要因が関係していると考えられる。

また，一方，簡易な評価者トレーニングを 2 回行った簡易グループの評価力は，全体的な評価力を上げることはできたが，「内容」を除くすべての評価項目に対する評価力を上げることができなかった。この結果から，簡易な評価者トレーニングの場合，回数を増やしたことで，学生の評価経験の浅さは改善できたかもしれないが，学生の評価力には影響を及ぼさないことがわかった。これらの結果から，一度の評価者トレーニングでは，学生の評価力に影響を及ぼすとは言えないが，トレーニングを二度以上行うことで，学生の評価力に伸びが見られたことから，評価者トレーニングは回数を重ねることで学生の評価力に影響を及ぼす可能性が高いことがわかった。つまり，評価者トレーニングを行うことにより，学生の評価力が向上し，学生による相互評価の信頼性が高くなる可能性があることがわかった。しかし，評価者トレーニングを授業時間内に実施するには，それなりの時間を要するため，教室内でより実現可能な評価者トレーニングができるようにするために方法を見直すべきであるという課題が残った。

第 7 章では，指導する評価項目別に評価者トレーニングを行うことで，評価者としての学生の評価力に影響を及ぼすかを調べた。相互評価を行う前の 3 週間をかけて授業時に評価項目別評価者トレーニングを行う 2 クラス（43 名）を詳細グループとし，相互評価を行う前週の授業時に 1 度のみ全評価項目に対する評価者トレーニングを行う 2 クラス（36 名）を簡易グループとした。筆者が評価項目を「発表」のみに絞り，さらに「姿勢」，「アイコンタクト」，「ジェスチャー」，「声の大きさ」，「明瞭さ」，「スライド」に細分化した。これに基づき，ルーブリックは変更したが，他は第 6 章で行った評価者トレーニングの実施方法と同じである。 1 回の評価項目別評価者トレーニングに要した時間は詳細グループにおいて約 30 分，一方，簡易グループのトレーニングに要した時間は約 45 分であった。

その結果，評価の合計点において，詳細グループ（r = .699），簡易グループ（r = .698）ともに中程度の相関が得られた。つまり，学生の総合的な評価力は，詳細グループ，簡易グループのどちらも高くないことが確認された。評価項目別にみると，中程度の相関から相関のない関係までさまざまであったことがわかった。つまり，評価項目別評価者トレーニングは，学生の評価力に影響を及ぼすとは言えない。

評価を行う直前に一度に全評価項目の評価者トレーニングを行うには，多くの時間を要し，また学生にとっても一度に覚えるものが多く，負担が大きいものと思われた。しかし，評価項目別に評価者トレーニングを行う方法は，1 回のトレーニングの所要時間を減らすことはできたが，結局各評価項目に対して 1 回ずつしか評価者トレーニングを行うことができず，学生の評価経験の浅さからくる評価力の無さがある程度補うことができたと思われるが，結果として全評価項目の評価力を上げることにはつながらなかった。実施方法を見直す必要性が今後の課題として残った。

第 8 章では，評価者トレーニングが評価者としての学生の評価力を高めるだけではなく，評価者トレーニングをプレゼンテーションの指導となり得るかを調べた。評価者トレーニングを行っていない 2015 年度の学生を「評価者トレーニングなしグループ」，評価者トレーニングを行った 2016 年度の学生を「評価者トレーニングありグループ」とした。評価者トレーニングが学生のプレゼンテーション力に影響を及ぼすかを調べるために，評価者トレーニングを行ったグループと，評価者トレーニングを行わなかったグループそれぞれの，後期における中間発表と最終発表の成績を比較し分析した。その結果，両グループの間における中間発表から最終発表への成績の伸びに差がなかったことから，評価者トレーニングが学生のプレゼンテーション力に影響を及ぼす可能性は低いことが考えられた。しかし，グループ別に分析した結果，評価者トレーニングを行わなかった場合，中間発表から最終発表への成績に変化は見られなかったが，評価者トレーニングを行った場合，最終発表の成績が中間発表の成績を若干ではあるが上回っていることことがわかった。このことは，評価者トレーニングは学生のプレゼンテーション力に少し影響する可能性があることを示している。つまり，評価者トレーニングは学生のプレゼンテーション力にわずかに影響を及ぼす可能性があることがわかった。また，評価者トレーニングの内容の違いが，学生のプレゼンテーション力に影響を及ぼすかを調べた結果，簡易的な評価者トレーニングを行う方が学生のプレゼンテーション力に影響を及ぼす可能性があることが確認された。

これらの結果から，評価者トレーニングが学生のプレゼンテーション力に影響を及ぼすためには，評価項目および評価者トレーニングの方法，実施回数，および実施する時期を見直し，手順を簡易にしたトレーニングを継続的に実施し，さらに発話トレーニングも併せることで，学生のプレゼンテーションの指導の一つとして活用できる可能性が示唆された。

第 9 章では，本研究であきらかになったことと，問題点と今後の課題についてまとめた。

本研究では，学生の相互評価の信頼性を調べるために，評価者としての学生の評価力に影響を及ぼすと考えられる 4 つの要因について調べた。その結果，学生の「プレゼンテーション力」，「英語力」そして「予備知識」も影響を及ぼさないことがわかったことから，学生の評価力の低さは，経験不足である可能性が高いことが考えられた。そして，評価者トレーニングを行い，学生の評価力が上がるかどうかを調べたところ，学生の評価力の無さはトレーニングを行うことである程度補うことが可能であることがわかった。さらに，評価者トレーニングは学生のプレゼンテーション力にわずかに影響を及ぼす可能性があることもわかった。

「プロジェクト発信型英語プログラム」は，学生が主体的に取り組めるプログラムのため，学生のモチベーションがとにかく最後まで落ちることがないことから，とて

も素晴らしい教授法の一つであると実感している。しかし，そこで行われている評価方法については，改善の余地があると筆者は考えている。具体的には，学生側においては，学生による相互評価の評価項目の見直し，および評価力のトレーニングの実施である。「プロジェクト発信型英語プログラム」では，学生一人一人が興味関心のあることに基づいてリサーチを行った成果を発表するため，すべての発表内容に対するスキーマを持っておくことは難しい。このため，学生には発表の「内容」ではなく，デリバリーに関する評価をさせ，教員はデリバリーの他に，内容や英語力も含めた評価を行うなど，学生と教員とで評価する項目を変えることも検討する必要がある。

一方，教員側においては，判定基準をプログラムの主旨に従って明確にしたルーブリックの作成，また担当教員間における評価者間信頼性の確保である。これらを行っていくことにより，このプログラムの到達目標が教員間においても，学生教員間においても明確になることで，どの教員が担当しても，同様のプレゼンテーションができ，さらにそこに学生のオリジナリティが加わった他に類を見ない発表ができる学生が増えていくことにつながっていくと思われる。

最後に，評価者トレーニングを行うことは，ただ単に教員による評価の一致の度合いをあげるためだけではないと筆者は考える。トレーニングを行うことにより，学生に評価力がついてくることから，クラスメイトの良い発表からも，そうでない発表からも学ぶことができるようになる。このように学生の批判的思考が高まることは，将来学生が自律した学習者になるために必要な判断力を身に着けることにつながる。また，評価者トレーニングを行い，学生にとって評価が難しい評価項目を知ることは，教員自身による評価をも見直す良い機会ともなり得る。つまりは，評価者トレーニングは学生の評価力をあげるためにだけでなく，教員に評価力をあげることにもつながると言えよう。

今後の課題として，本研究では，評価者としての学生の評価力に影響を及ぼす要因を明確に突き止めるには至らなかったことから，その要因を探るべく，引き続き研究を続ける必要がある。また，二回にわたる評価者トレーニングを行っても，発表の内容に関する項目に対する学生の評価力を上げることができなかったことから，今後も，これらの評価項目に対して何が影響を及ぼすのかを探求する必要がある。そして，本研究で明らかになった評価者トレーニングの改善点を見直し，手順を簡易にしたトレーニングを継続的に実施することで，学生の評価力を高められるかどうかを検証する必要がある。最後に，評価者トレーニングが，学生のプレゼンテーションの指導の一つとして活用できるようにするために，発話トレーニングも併せることで，学生のプレゼンテーション力を向上させることができるかどうかを検証したい。

謝辞

この博士論文の執筆にあたり，、実に多くの方々にお世話になりました。この場をお借りして、感謝の意を述べさせていただきたいと思います。

まず，鈴木寿一先生には，6 年間という長い間，本当に懇切丁寧なご指導をしていただきました。研究に向かう姿勢や研究に関する困難克服のための具体的な方策まで丁寧に教えていただきました．その温かく優しい励ましと，時に厳しいご指導がなければ，本論文を完成させることはできませんでした。心より感謝致しております。また，ご多忙の中，副査を引き受けてくださいました，筑波大学の平井明代教授には，貴重なご指導とご助言を賜りました。感謝申し上げます。同じく副査としてご助言を下さいました京都外国語大学の相川真佐夫教授，第 1 次発表，第 2 次発表の審査をしてくださいました，吉田真美教授，安木真一教授にも，本論文の展開に関わり，貴重なご教示と優しい励ましをいただきました。特に，吉田真美教授には，子育てと研究を両立させる難しさから，幾度となく博士論文の執筆をあきらめかけた筆者に，同年代の子供を持つ母として，研究の進め方に関する具体的なご助言を多く賜りました。このような先生方の励ましは，時にくじけそうになる長期にわたる研究の日々の支えとなりました。深く感謝します。

筆者の前勤務校である，立命館大学　生命科学部の山中司教授には，ご多忙な中，筆者の博士後期課程進学に際し，多くのご助言と励ましをいただきました。また，「プロジェクト発信型プログラム」の同僚の先生方にも大変お世話になりました。ありがとうございました。

学会発表の折に貴重な質問やご意見をくださった，高校・大学の諸先生方，また，学会誌の査読委員の皆様，有益で貴重なご助言を本当にありがとうございました。

京都外国語大学大学院生の皆様にはとても親しくしていただきました。また現職教員として教育と研究を両立しておられる姿は，筆者の論文執筆の大きな励みとなりました。本当にありがとうございました。

筆者の研究の主旨を理解し，快く参加してくれた学生達に深く感謝します。彼らの明るい笑顔と優しさにいつも励まされておりました。

最後に，筆者の仕事，研究，家庭の両立を常に励まし，支え続けてくれた家族に心から感謝します。家族の支えがなければ，この論文は完成していません。何かあれば，いかなる時も遠方より駆けつけてくれて，筆者に代わり家事・育児をしてくれた母，執筆の邪魔をしないように気を使いながら，応援メッセージを手紙に書いて，そっと筆者の机に置きに来てくれた娘，そして全国学会誌に筆者が投稿した論文が掲載されることが決まった際，誰よりも喜んでくれて，一緒に祝ってくれた主人に感謝の気持ちでいっぱいです。特に主人は，筆者が博士後期課程進学を決めたときから，博士論文を執筆し終えるまで，一番の理解者として，ずっと筆者を支えてくれました。深く感謝します。

本論文は，このように多くの皆様に支えられて完成できたものであると，改めて実感

しております。ここに厚く感謝の意を表します。

第1章　はじめに

1.　研究の背景・動機

近年の英語教育においては，コミュニケーション重視の傾向にあり，多くの教員による様々な取り組みのもと，教室内においても学生が積極的にコミュニケーションする姿が見受けられる。

しかし，スピーキングの評価に関してはあまり適切に行われていない。学生のスピーキング･パフォーマンスを評価するのは,担当教員が一人で行うのが現状であろう。しかもその評価とは，あくまで成績をつけるための評価であって，日を置いて再度評価し，評価者内信頼性を確保していないことが多い。そのため，一人の評価者による評価は，信頼性があるとは言い難い。また，実践面においても，学生一人ひとりの評価を担当教員一人で行うのには，相当な時間と労力が必要とされる。

評価に対するこれらの問題点の解決策の一つとして，学生による相互評価の活用が考えられる。相互評価とは，個々の学生のパフォーマンスを学生同士が相互に評価を行うことであり，教員による評価を補足するために，また学習者に学習目標を明確にするために，スピーキングのクラスにおいて，ますます取り入れられる傾向にある。学生による相互評価を活用できれば，いろいろな視点から，より客観的に学生のスピーキングを評価することが可能になることが考えられる。また実践面においても，学生による相互評価は授業時間内に行うことができるため，教員の評価にかかる時間や労力が少しでも軽減されるとも考えられる。

鈴木（2012）により開発、実践された「プロジェクト発信型英語プログラム」は、発信型の英語能力を育成するため、学生は自身の興味・関心に基づいた内容を英語のプロジェクトとして発信している。授業では学生同士が英語でコミュニケーションを行っており,今でも見られるような教員から学生への一方向の授業ではない。しかし，その評価においては，教員一人で行っている。教員たった一人の評価では，ややもすれば偏った評価になりかねない。実際に，学生の発表においても，教員からすれば，よく理解できないような内容であっても，学生の反応がとてもよく，発表内容をよく理解して，熱心に話に聞き入っている学生が時折ではあるが見受けられた。そこで，学生の視点も組み入れることで，様々な観点からより客観的に学生の発表内容の評価をできないだろうかと考えたのが，本研究の出発点である。

しかし,学生による相互評価を取り入れるには,その信頼性を検証する必要があり，これまで多くの教員や研究者により様々な研究が行われてきた。その結果，学生による相互評価は信頼できるとする説（Fukazawa, 2010; Nakamura, 2002; Okuda & Otsu, 2010）と，信頼できないとする説（Freeman, 1995; 笠巻, 2016）と，相反する見解が見受けられた。しかし，信頼できないとする場合，何が原因なのかを検証した報告は筆者が調べた限りでは見当たらなかったので，その点を調べることにした。

2. 本研究の構成

第 1 章では，最近の日本の大学でのスピーキング・クラスにおいて行われている評価方法，またその問題点，本研究を行った背景および動機について述べている。

第 2 章では，スピーキングの評価の観点，評価方法，評価に影響を及ぼす要因についての先行研究を紹介する。また，学生による相互評価を活用するにあたり，相互評価の利点および欠点についての先行研究と，相互評価の信頼性を検証した実証的研究をまとめ，これまでの研究で明らかになってきたことと今後行うべき研究の課題を明らかにする。

第 3 章では，評価者としての学生の評価力に影響を及ぼすと考えられる 4 つの要因，学生の「プレゼンテーション力」，「英語力」，「予備知識」そして「評価者トレーニングの有無」のうち，学生の「プレゼンテーション力」が評価者としての学生の評価力に影響を及ぼすかについて検証する。また，「プレゼンテーション力」別に見た，学生の評価傾向についても報告する。

第 4 章では，上述の要因のうち，学生の「英語力」が評価者としての学生の評価力に影響を及ぼすかについて検証する。また，英語力別に見た学生の評価傾向についても報告する。

第 5 章では，上述の要因のうち，プレゼンテーションの内容に対する学生の「予備知識」が評価者としての学生の評価力に影響を及ぼすかについて検証する。また，予備知識の多少別に見た学生の評価傾向についても報告する。

第 6 章では，上述の要因のうち，「評価者トレーニングの有無」が評価者としての学生の評価力に影響を及ぼすかについて検証する。また，評価者トレーニングの回数，内容や質の違いが学生の評価力の向上に影響を及ぼすかについても報告する。さらに，学生の評価力を高めるためには，評価者トレーニングをどのように活用することが有効なのかについても考察する。

第 7 章では，第 6 章で行った評価者トレーニングを，到達目標となる評価項目別にトレーニングを行い，実際の相互評価の直前ではなく，評価者トレーニングを授業の一部に組み入れることで，学生の評価力に影響を及ぼすかについて検証する。

第 8 章では，評価者トレーニングを受けた学生と，受けなかった学生それぞれのプレゼンテーション力の比較を行い，評価者トレーニングがプレゼンテーション力向上のための指導方法の一つとなり得るか，その有効性について報告する。さらに，第 7 章で行った内容や質が異なった評価者トレーニングが，学生のプレゼンテーション力に与えた影響についても報告する。

第 9 章では，第 3 章から第 8 章までの 6 つの研究から得られた結果をまとめ，それらの研究の問題点や限界点を述べたのち，今後の研究の方向性を示す。

最後に，資料として実証研究で用いた評価シートを添付する。

第2章　先行研究

1.　スピーキングの評価

1.1　評価の観点

言語テストの分野において, スピーキングの評価は最も難しいと言われている(馬場, 1997; Fulcher, 2003; Underhill, 1987)。その理由として，評価の観点について、正確さ，適切さ，自主性，繰り返し，伝達度，情報の量，内容の正確さ，スピード，つなぎ言葉の用い方，談話としての整合性，発話しようとする意欲，わかりやすさ，流暢さ，発話量，統語的複雑さ，柔軟さなど様々な観点があることが挙げられる（馬場, 1997：岡, 1984)。また，人によって「スピーキングとは何か」に対する考え方が違うこと(馬場, 1997)，評価の観点が人によって異なること，そして評価が主観的になりやすくなること（Hughes, 1989; McNamara, 2000）もスピーキングの評価の難しさとして挙げられる（Douglas, 1994)。

1.2　スピーキングの評価方法

スピーキング力を客観的に評価する方法として，既存の英語運用能力テストでは、主に英語に関する知識、発音、流暢さなどの言語側面を評価して英語能力を測るものとして、TOEFL, TOEIC, 英検、Versant など、これまで数多く開発されている。

話せるかどうかををテストするには，実際に話させることが必要であり（石川他, 2011; 馬場, 1997)，スピーキング力を測定するには試験者と被験者が一対一で実際パフォーマンスを行うことが最も適しているとされている（Brown, 1993)。

ACTFL(全米外国語教育協会)が開発した OPI (Oral Proficiency Interview) という、「機能・タスク」、「場面・内容」、「テキストの型」、「正確さ」の 4 つの要素をガイドラインに沿って、全体的・総合的に判断するインタビューテストがある。筆者自身も、ACTFL (全米外国語教育協会) と㈱アルクが共同で開発した、インタビュー形式でスピーキング力を測定する SST (Standard Speaking Test) および T-SST (Telephone Standard Speaking Test) の試験管・評価官の資格の取得に積極的に取り組み、英語の発話能力を言語的側面からのみでなく包括的に評価する方法を学び、多くの被験者に実施してきた経験を持つ。

しかし、試験者と被験者が行うパフォーマンスは、あくまでテストという枠組みで、話す内容をある程度限定してしまっている。この点について、Norman (1998) は、学習者の口頭運用能力を直接評価するには、コミュニケーション活動における実際のパフォーマンスを評価する必要があると主張している。

1.3　スピーキングの「内容」の評価

コミュニケーションの手段としての英語とは、自分の意見や考え、情報やメッセージを伝えるためのものであり、コミュニケーションの中でなされるスピーキング力においては、英語の発音、流暢さ、および正確さといった言語側面よりも、むしろ話す内容やその伝え方が重視されるべきである。伝えたいメッセージが伝わること、つまりスピーキングの内容が評価されるべきである。

Sato（2011）は、実験の参加者が伝えようとしている考えの質を、ライティングの試験同様、モノローグのパフォーマンスを含むオーラルのテストにおいても評価基準の一つに入れるべきであると主張している。そして、言語的側面に限定して評価することは、第２言語学習者のコミュニケーション能力を測ることにおいて、間違った評価をしかねないとも指摘している。

しかし，スピーキングの内容に対する評価においては、もう 30 年以上も前からその必要性が指摘されているが、いまだに明確な評価規準がなく、多くの研究者や教員たちが模索し続けている（馬場, 1997）。

このコミュニケーションの「内容」が積極的に評価されていない原因の一つとして、数値では測れない、あるいは評価すること自体が非常に難しいため、いわゆる「発信」を評価するための取り組みを報告したものは少ない。

またスピーキングの「内容」の評価に学生相互評価を取り入れる試みの報告は，筆者が調べた限りでは見当たらなかった。

1.4　評価に影響を及ぼす要因

評価には，評価者の様々な個人的な背景が評価に影響を及ぼす（Taylor & Galaczi, 2011）と言われているが，具体的に評価者の何が，評価および評価の信頼性に影響を及ぼしているかを検証した研究は少ない。

Tanaka (2017)は，大学 1 年生を対象に，英語によるプレゼンテーションに対して学生に相互評価をさせ，評価者としての学生の性格によるバイアスの有無を検証した。その結果，他者との関係を重視する学生の評価は甘くなり，他者からの評価を気にする学生の評価は厳しくなることが明らかとなった。このことから，学生による相互評価には，評価者としての学生の性格が影響を及ぼす可能性があるとしている。

また，De Grez (2010) は，評価者の性格，特に「自己効力感」が評価に影響すると報告している。

上に述べたように, 評価者個人の性格が評価に影響を及ぼす可能性があることから，一人の評価者による評価は，たとえ十分なトレーニングを積んだとしても，信頼性があるとは言い難い（Hughes, 1989 ; Underhill, 1987）。

2. 相互評価

2.1 相互評価の利点

相互評価の利点として, 学生の授業参加を促すことができる（Brown, 1998; Nakamura, 2002; 三木・笠巻, 2017），学習者に学習目標を明確にさせる（Goh & Burns, 2012; Fukazawa, 2010; Luoma, 2004），学習者の学習に対する動機づけが高まる（Nakamura, 2002），そして，相互評価を通して，学生がお互いに学び合うことができる（Luoma, 2004; 藤原他, 2007b）といった点が挙げられる。さらに，他の学習者を評価することが自己への振り返りにもなり得るといった相互評価の有効性も報告されている。(菅沼, 2013; Yang et al., 2006)。

また，スピーチ・プレゼンテーションのクラスで，実際の評価に用いられる評価項目を使って相互評価をさせることは，学生自身がプレゼンテーションをする際に，その評価項目を到達目標として準備することにも役立ち（Nakamura, 2002），発表をしている間，評価項目に関わるスキルに集中することができる（Brown, 1998）といった効果も報告されている。

さらに，Goh & Burns（2012）は，学生に相互評価をさせることにより，評価がどのように行われるのかを学生に意識させることで，自らの学習により責任を持たせ，自立した学習者へと導くことができるとしている。

また，三木・笠巻（2017）は，学生による相互評価を可視化することが，学生の心理面に影響を及ぼすかを調べた。クラスメイトであるオーディエンスがグループで行うポスター・プレゼンテーションを良いと判断した場合には，シール台帳にシールを貼っていくという試みを行った。授業後のアンケート調査により，相互評価を可視化することは，学生の心理面にネガティブな影響を与えることはなく，むしろ学生の学習意欲を高めた可能性があると報告している。

また，相互評価を教員による評価に補足的に活用することによって，評価にかかる時間や労力といった教員の負担を軽減できることも利点として挙げられる。（Brown, 1998; Fukazawa, 2010; Luoma, 2004; Okuda & Otsu, 2010）。

2.2 相互評価の欠点

一方，相互評価の欠点として，学生相互評価は，教員による評価よりも若干甘くなる傾向がみられる点と，学生は極端な評価を避け，評価の幅が狭いことが指摘されている (Cheng & Warren, 2005; Hughes & Large, 1993; Fukazawa, 2010; 笠巻, 2016; Otoshi & Hefferman, 2007)。

また，Brown (1998)は，評価が主観的になりやすいため，信頼性に欠くと指摘している。

Cheng & Warren (2005) は，大学生を対象に，学生の英語力の評価に対する態度と，

他の評価項目（準備，内容，構成，発表の仕方）の評価に対する態度を比較した。その結果, 評価自体にはそれほど大きな差は現れなかったが, 他の評価項目と比べると，英語力の評価に対して，あまり評価をしたがらないということがわかった。その理由として，学生が自分の英語力がクラスメイトの英語力を評価するのに十分ではないと感じていることを挙げている。

この点において，Luoma（2004）も，学生に言語面を評価させることは適当ではないとしており，むしろ，授業内で学生に課すタスクに関連した評価をさせるべきであると主張している。

また，藤原他(2007a)は，学生による相互評価では，相手に高い評価をすることで，自分に対しても高い評価をしてもらいたいという気持ちから，クラスメイトに対して否定的なコメントを抑える現象が起こることを指摘し，これを「お互い様効果」と名付けている。

上に述べたように，相互評価に対する不安や，友達を評価することに対する気まずさといった心理的な問題点も報告されており（Hanarahan & Issacs, 2001），学生の評価に対する心理的な要因が評価に影響を及ぼしかねないことも欠点の一つとして考えられる。

2.3 相互評価の信頼性

学生相互評価の信頼性には，信頼できるとする説と，信頼性に欠けるとする説と，2つの相反する見解が見受けられる。

Nakamura（2002）は，大学のオーラル・プレゼンテーションクラスの評価方法として，教員による評価と学生による相互評価の両方を取り入れ，項目応答理論のラッシュ･モデル[1]の FACET 理論を適用してデータ分析を行った。5 名の学生評価者のうち，1 名のみがミスフィットした評価者であると判断されたが，他の 4 名においては，評価の厳しさには差があるが，一貫した評価を行っていたことから，学生は信頼性を備えた評価者になりうると主張している。

Fukazawa（2010）は，高校生を対象に，授業 1 時間をかけて，各レベルのスピーチの例のビデオを用いた評価者訓練を行った。ラッシュ・モデル分析を用いて，スピーチ発表における相互評価を検証した結果，大半の学生はミスフィットしていないことがわかった。また，学生による相互評価と教員による評価の相関を調べたところ，高い相関（r = .79～.93）が得られたことから，高校生においても，学生による相互評価は信頼できるとし，実際の成績の一部として活用することが可能であると主張している。

また，Okuda & Otsu（2010）も，日本人大学生を対象に，実際の評価の前に，教員が実演しながら評価項目を説明し，それを評価の度に繰り返すという評価者訓練を行

った。教員による評価と学生による相互評価の一致の度合いを相関を用いて調べた結果，両者の間に高い相関（r = .82）が得られ，学生による相互評価を最終評価に組み入れることができるとしている。

Luoma（2004）も，相互評価は教員による評価に取って代わることはできないが，補足的に使うことはできるとしている。

このように，学生相互評価の信頼性について肯定的な研究がある一方で，否定的な見解もある。Freeman（1995）は，大学生を対象に，グループ・プレゼンテーションの評価，特にプレゼンテーションの内容と発表の仕方をグループで評価させ，教員による評価との相関を調べた。その結果，教員による評価と学生による相互評価には中程度の相関が得られた。しかし，学生がとても良いプレゼンテーションには低い評価を，逆にあまり良くないプレゼンテーションには高い評価をしてしまう傾向が見られたと報告している。学生による相互評価の信頼性や評価の一貫性は，学生の評価者としての評価経験の有無と関係していることが，Weigle（1994），山西（2004）の研究でもすでに述べられている。

Oi (2012) は，高校生を対象に，クラスメイトの英語のスピーチに対して，自己評価と相互評価を 1 か月間の間にそれぞれ 3 回ずつ行い，教員による評価との相関が変わるかを調べた。その結果，相互評価においては，「内容」，「発表」「言語使用」の評価項目のうち，「発表」に対してのみ，教員による評価との間に中程度の相関が見られたが，他の項目においては相関が見られなかった。その理由として，学生は，クラスメイトのスピーチを評価することに対して自信がなく，また心理的に負担を感じていることがわかった。実施した 3 回の相互評価のいずれにおいても，相関が見られなかったことから，学生による相互評価の信頼性は低いとしている。

笠巻（2016）では，大学 1 年生 61 名を対象とし，学生が行った英語による口頭発表に対する学生による相互評価と教員による評価を調べることにより，学生による相互評価の信頼性を検証した。そして、相関が高い場合、低い場合それぞれにおいて、その背景にある原因および理由を考察し、スピーキングの内容を評価するにあたり、学生の視点を取り入れられるかどうかを検討した。

その結果，学生一人ひとりに対する総合評価においては、相関係数は、r = .70 で、教員評価と相互評価の間には比較的高い相関があることがわかった。この結果により、総合評価における教員の評価と学生の評価の一致の度合いは高いと言える。しかし，項目別に見ると，準備と発表の仕方においては，相関係数は r = .70 で、教員による評価と学生による相互評価の間には比較的高い相関があると考えられるが，リサーチ（r = .57）とオリジナリティ（r = .56）においては，教員による評価と学生による評価の相関は中程度であることから，これら二つの項目においては，教員による評価と学生による評価の一致の度合いは高いとは言えない。つまり，総合評価における教員の評

価と学生の評価の一致の度合いは高くても，評価項目によって教員による評価と学生による評価の一致の度合いが異なることがわかった。

また，学生による相互評価の平均値が、全ての評価項目、総合評価ともに一貫して教員より高い傾向にあり，標準偏差においては、全ての評価項目、総合評価ともに一貫して教員より小さい傾向が見られた。このことから，学生による評価は，教員より甘くなり、極端な評価を避けるという２つの傾向があることもわかった。さらに，リサーチとオリジナリティといういわゆる発表内容に関わる評価項目に対して、これらの傾向が顕著に現れた。

さらに，相互評価の実施後に，学生に自分が良いと思うプレゼンテーションとはについてアンケート調査を行ったところ，、学生の回答は，主に発表の仕方の上手下手に集中していた。理由を尋ねたところ、「発表の仕方が良くないと何を話しているのかよくわからず、肝心の内容が伝わらなくなってしまう」というものであった。発表の「内容」に関しては、特に学生から明確な回答が得られていない。これらの結果から、学生が何を持って「内容」の良し悪しを判断するのかがわかっていなかったものと考えられる。一方、発表の仕方にはついては、学生はその良し悪しをほぼ的確に理解していると思われる。このことは、教員による評価と学生による相互評価が比較的高い相関を示したこととも合致する。

これらの結果から，学生相互評価の信頼性は低いと判断された。そのため、スピーキングの「内容」の評価には，学生による相互評価を補足的に活用することは現段階では難しいということが明らかになった。

また，評価者となる学生がクラスメイトのスピーキング・パフォーマンスの評価を匿名で行うことで，相互評価の信頼性を高めることができるかについて検証した研究がある。岡田（2017）は，日本人大学生 25 名を対象に，クラスメイトのスピーキング・パフォーマンスを匿名，また記名の両方で相互評価をさせ，得点をつけさせる方法とコメントを記入させる方法の両方を行うことで，評価の信頼性の違いを検証した。その結果，評価得点においては，評価者名を記入した相互評価は，匿名による相互評価より低い評価になることがわかった。つまり，匿名で相互評価を行うより，記名で行う方がより厳しめの評価になる傾向があることが明らかとなった。一方，自由記述によるコメントにおいては，匿名による相互評価は，学生にとってコメントを記入しやすく，また評価を厳しくしやすいといった長所がある一方で，学生による記述コメントに対する責任の所在が明確にならない点を短所として指摘している。結果として，匿名による相互評価と，記名による相互評価のどちらが信頼性が高いかについては明らかにならなかったが，学生同士による人間関係が評価に影響を及ぼしい兼ねない相互評価を，教育の現場に導入するためには，クラスメイトのスピーキング・パフォーマンスの評価に対して，責任のある評価を与えることに注意を払う必要があると主張

している。

一方，Hirai *et al* (2011)は，学生評価者を匿名にすることと，実際の評価を行う前に，クラスメイトとペアになり，クラスメイトのオーラル・パフォーマンスについてディスカッションをすることで，日本人の学生による相互評価が教員による評価とどの程度相関するのかを調べた。その結果，学生による相互評価の信頼性を高めるために評価者を匿名にすることは重要な要因になることが明らかになった。しかし，評価の前にペアでディスカッションを行うことは，学生が一人で行う評価と比べて，必ずしも信頼性を高めるものではないことを明らかにしている。また，教員による評価と学生による相互評価の相関は高いものから低いものまであり，学生による相互評価は教員による評価に取って代わることができる信頼性の高いものではないと主張している。そして，相互評価の意義として，学生による相互評価を教員による評価に近い信頼性の高い評価として求めるのではなく，クラスメイトを評価することよりもむしろフィードバックを与えるために相互評価を使うことで，評価自体が協同学習としてより有益なものになる可能性があるとしている。

注

1. 項目応答理論とは「ある困難度を持ったテストの項目に，ある能力を持った受験者がどのように応答するか，ということに確率的なモデルを設定し，それに基づいて受験者の応答データを分析したり，テストを開発するための理論」（静・竹内・吉澤，2002, p.100）であり，項目応答理論の 1 パラメータ・ロジスティック・モデルが，ラッシュ・モデルと呼ばれている。

第 3 章　研究 1：学生の「プレゼンテーション力」が評価者としての学生の評価力に影響を及ぼすか？

1.　はじめに

笠巻（2016）は，教員による評価と学生による相互評価の相関を調べ，相関が高い場合，低い場合において，その背景にある原因および理由を考察し，スピーキングの内容を評価するにあたり，学生による相互評価の信頼性を検証し，学生の視点を評価の一部に取り入れられるかを検討した。教員による評価と学生による相互評価との間には，中程度の相関が得られたが，学生の発信する内容を評価する項目においては，さらに相関が低くなる傾向が認められたことから，学生による相互評価の信頼性は低く，評価の一部に補足的に組み入れることは難しいということがわかった。学生の評価に影響を及ぼす要因として，学生のプレゼンテーション力，英語力，予備知識，および評価者トレーニングの有無が考えられた。

2.　研究1の目的

研究 1 では，評価には上述の 4 つの要因の一つである，学生のプレゼンテーション力の違いによって，プレゼンテーションに対する学生による相互評価と教員による評価との相関が変わるかを調べ，学生のプレゼンテーション力が学生の評価力に影響を及ぼすかを検証すること，また，その背景となる学生の評価傾向およびその原因を調べることを目的に，分析 1 と分析 2 を行なった。

分析 1. 学生のプレゼンテーション力は学生の評価力に影響を及ぼすか。
分析 2. 何が学生の評価傾向に影響を及ぼすか。

3.　研究1における指導と評価方法

3.1　指導

「プロジェクト発信型英語プログラム」では，各自が取り組むプロジェクト活動を通して，リサーチ，ディスカッション，プレゼンテーションのスキルを学び，また，自分自身のプロジェクトの成果を英語で発信する力を身につけることを到達目標としており，教員は，学生がプロジェクトを進めるにあたり必要なアドバイスをするファシリテーターに徹する（鈴木，2012）。

筆者が担当した 1 年生の前期では，教科書 [2] に沿って，プロジェクト，リサーチ，プレゼンテーション・スキルの基本を学び，簡単なプロジェクトを行ってセルフ・アピールをした。鈴木（2012）に基づいて主な授業の流れを表 1 にまとめた。学生は毎回の授業において，自分が選んだテーマに基づいて取り組むプロジェクトの進捗状況

についてグループ・ディスカッションを行い，その後，その結果をクラス内で発表し意見交換をした。教員はその都度必要な助言を与えた。そして学生は，各自が取り組んだプロジェクトの成果について中間発表（5 分間）と最終発表（8 分間）を行った。

表 1
主な授業の流れ

	内容
クラス外ワーク	前回の授業で学んだことを活かして，授業外でプロジェクト活動を行い，その報告を英文で書き，学習管理システムの manaba に授業前に提出する。
Step 1	クラス外ワークで進めてきた自分のプロジェクト活動について 3〜4 人のグループ内で発表し合い，意見交換をする。その際，必ず moderator を一人選び，時間を考慮に入れながら進める。
Step 2	各グループの中から 1 人を選出し，自分のプロジェクトについて，またグループ内で話し合ったことについて，クラス全体に対して発表し意見交換をする。その後，教員がその発表についてフィードバックを与える。
Step 3	必要なスキルや表現を学んで練習し，それらを活用してプロジェクトを進める。

3.2 評価方法

「プロジェクト発信型英語プログラム」では，学生の発表の際に，教員による評価のほかに，オーディエンスとなる学生を積極的な聞き手にする目的で，学生による相互評価を取り入れており，教員と発表者以外の学生は，同じ評価シート（資料 1 参照）を用いて，発表者のプレゼンテーションを聴くのと同時に評価を行う。この評価シートは，学生は学習管理システムである manaba よりダウンロードして使用する（三木・笠巻，2017）。

評価項目は，到達目標を鑑みて作成された準備，リサーチ（内容と構成），オリジナリティ（熱意と説得力），発表（発音，声の大きさ，速度，アイコンタクト，明瞭さ）の 4 つで，判定基準は，5 = excellent，4 = good，3 = so-so，2 = poor，1 = inadequate の 5 段階で評価を行う。

学生は，発表者全員の発表終了後にその評価シートを manaba に提出する。その際，発表者に気兼ねすることなく，率直な評価ができるようにとの配慮から，学生同士は閲覧ができず，提出者本人と教員だけが見ることができるように設定されている。また，学生には，この評価シートは提出させるのみで成績に加味しないため，素直につ

けるように指示している。

このプログラムでは，この評価シート，評価項目，判定基準について，発表前週の授業時に学生に口頭で説明するのみで，事前に評価者トレーニングなどは行っていない。

4. 研究方法

4.1 参加者[3]

研究 1 の参加者は，滋賀県内の私立大学における筆者担当の大学 1 年生 4 クラス計 61 名のうち，学生一人ひとりのパフォーマンスに対する教員による評価と学生による評価の相関係数が何らかの事情で欠けている学生[4]と，相関係数が外れ値となった学生 19 名を除いて，42 名を分析対象とした。参加者の TOEIC IP テストの平均点はおよそ 500 点である[5]。

4.2 手続き

学生はクラス全員の前で，プロジェクトの成果の中間発表として，一人 5 分間ずつ英語でプレゼンテーションを行い，3.2 で述べた評価方法でこの中間発表の評価を行った。ただし，本研究では，判定基準については，評価に迷った場合のために中間点を設けたため，原則的には 5 段階評価のところを，実質的に 10 段階評価（10 = excellent, 8 = good，6 = so-so，4 = poor，2 = inadequate）とした[6]。発表終了後，manaba に提出された学生の評価シートを集計し，教員による評価と学生による相互評価を分析の対象とした。

また，本研究では，教員の評価者は筆者一人である[7]。そのため，評価者内信頼性を調べる必要があり，1 回目の評価の約 4 ヶ月後に，学生 15 名を抽出し，筆者が録画しておいた学生の発表ビデオを用いて，後日再評価を行った。各評価結果の相関係数を求め，評価者内信頼性係数を算出した（r = .90）。

4.3 分析 1: 学生のプレゼンテーション力は学生の評価力に影響を及ぼすか。

4.3.1 分析方法

4.3.1.1 学生の相互評価と教員による評価の相関係数の算出方法

まず，各学生のプレゼンテーションに対する一人ひとりの学生による評価点と教員による評価点の相関を算出した。次に，全学生のプレゼンテーションに対する学生と教員の評価の相関係数をを合計して平均した。

4.3.1.2 得られたのデータの分析方法

教員は優れた発表には高い評価をつけ，あまり良くない発表には低い評価をしてい

るが，学生が正しく評価できるかどうかを調べるために，学生もまた教員による評価と同じように，優れた発表には高い評価を，あまりよくない発表には低い評価をしているかを調べた。筆者が各学生のプレゼンテーションを採点して得られた得点を偏差値化し，偏差値 55 以上を上位群 (13 名)，偏差値 45 から 54 を中位群(13 名)，偏差値 45 未満を下位群(16 名)とし，学生が行ったプレゼンテーションに対する学生による相互評価と，教員による評価との相関係数を成績群ごとに算出した。学生一人ひとりのパフォーマンスに対する学生による評価の合計点の正規性の検定を行ったところ，表 2 が示す結果になった。

正規性が認められなかったことと（上位群 : D (215) = 0.12, p <. 05，中位群 : D (202) = 0.16, p < .05，下位群 : D (240) = 0.12, p <. 05），担当クラスの数の関係で，データ数が少ないことにより，スピアマンの相関係数を用いた。次に，プレゼンテーション力の上位群，中位群，下位群の間に差があるかについてクラスカル・ウォリス検定を用いて調べ，その後，マン・ホイットニーの U検定を用いて多重比較をおこない各 2 群間の相関係数の差を調べ、効果量を算出した。

表 2

正規性の検定結果

	Kolmogorov-Smirnov(a)		
	統計量	自由度	有意確率
上位群	.118	215	.000
中位群	.163	202	.000
下位群	.122	240	.000

5. 結果と考察

5.1 群別の分析結果と考察

5.1.1 上位群

プレゼンテーション力の成績上位群における，学生による相互評価と教員による評価との相関係数の記述統計量を表 3 に示す。図 1 は，上位群の箱ひげ図である。

表 3

上位群の記述統計量　（*n* =13）

	M	*SD*
準備	.370	.333
リサーチ	.338	.183
オリジナリティ	.326	.192
発表	.469	.166
合計	.529	.184

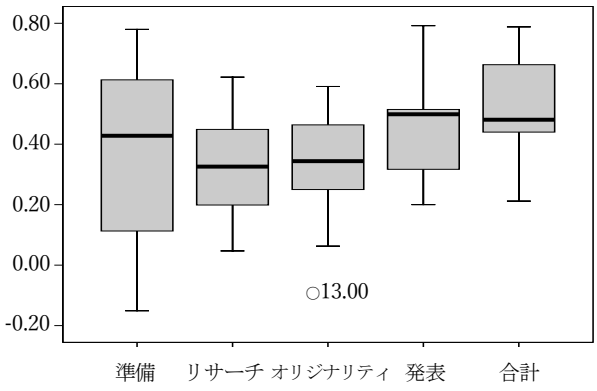


図 1．上位群の箱ひげ図

上位群においては，評価の合計点において中程度の相関が得られた（*r* = .529）。しかし，評価項目別に見ると，発表（*r* = .469）においては中程度の相関が得られたが，準備（*r* = .370），リサーチ（*r* = .338），オリジナリティ（*r* = .326）においてはそれぞれ弱い相関が得られた。また，図 1 からは，他の評価項目と比べると，特に標準偏差の大きい評価項目である「準備」（*SD* = .333）は，相関の程度が中程度（*r* = .703）から相関がない関係（*r* = .037）になるほど，相関係数がばらついていることがわかった。これらの結果から，上位群の学生の評価と教員による評価の相関は高くないことがわかった。学生による評価と教員による評価の間に高い相関がないことは，学生の評価力は高くないと言える。つまり，上位群の学生のプレゼンテーション力は評価能力とは関係ないと言えよう。

5.1.2　中位群

プレゼンテーション力の成績中位群における，学生による相互評価と教師による評価との相関係数の記述統計量を表 4 に示す。図 2 は,中位群の箱ひげ図である。

表 4

中位群の記述統計量（$n = 13$）

	M	*SD*
準備	.472	.071
リサーチ	.206	.399
オリジナリティ	.333	.156
発表	.587	.227
合計	.580	.186

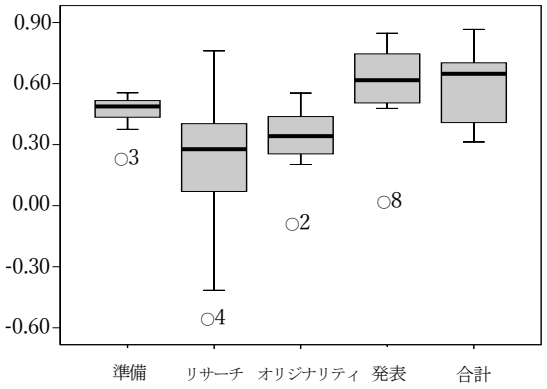


図 2. 中位群の箱ひげ図

中位群においては，教師による評価との相関は，評価の合計点において，中程度の相関が得られた（$r = .580$）。しかし，評価項目別に見ると，準備（$r = .472$）と発表（$r = .587$）においてはそれぞれ中程度の相関が得られたが，リサーチ（$r = .206$），オリジナリティ（$r = .333$）においては，それぞれ弱い相関が得られた。また，図 2 からは，標準偏差の小さい「準備」（$SD = .071$）においては，相関の程度が中程度（$r = .401$～$.543$）に留まり，ばらつきは見られないが，特に標準偏差の大きい「リサーチ」（$SD = .399$）においては，相関の程度が中程度（$r = .605$）からマイナスの関係（$r = -.193$）までと 3 段階におよび，相関係数のばらつきがとても大きいことがわかった。これらの結果から，中位群の学生の評価と教員による評価の相関は高くないことがわかった。学生による評価と教員による評価の間に高い相関がないことは，学生の評価力は高くないと言える。つまり，中位群の学生の評価力も高くないと言えよう。

5.1.3 下位群

プレゼンテーション力の成績下位群における，学生による相互評価と教師による評価との相関係数の記述統計量を表 5 に示す。図 3 は，下位群の箱ひげ図である。

表 5

下位群の記述統計量（$n = 16$）

	M	*SD*
準備	.496	.224
リサーチ	.315	.243
オリジナリティ	.297	.285
発表	.497	.183
合計	.609	.198

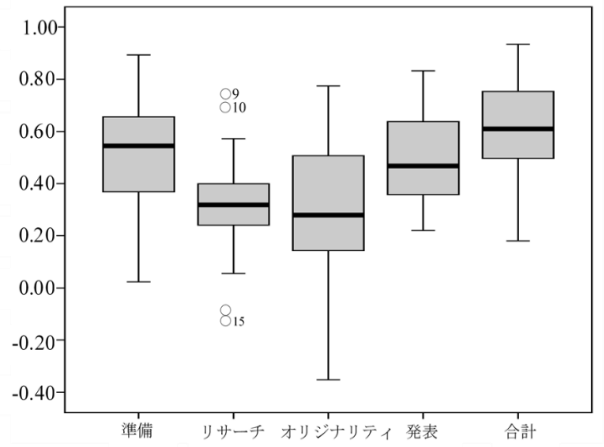


図 3. 下位群の箱ひげ図

下位群においては，教師による評価との相関は，評価の合計点において，中程度の相関が得られた（$r = .609$）。しかし，評価項目別に見ると，準備（$r = .496$）と発表（$r = .497$）においてはそれぞれ中程度の相関が得られたが，リサーチ（$r = .315$），オリジナリティ（$r = .297$）においては，それぞれ弱い相関が得られた。また，図 3 からは，「準備」（$SD = .224$）は弱い相関（$r = .272$）から高い相関（$r = .720$），「リサーチ」（$SD = .243$）は中程度（$r = .558$）から相関がない（$r = .072$），「オリジナリティ」（$SD = .285$）は同じく中程度（$r = .582$）から相関がない（$r = .012$）と，それぞれ相関の程度が 3 段階に大きくばらついていることがわかる。これらの結果から，下位群の学生の評価と教員による評価の相関は高くないことがわかった。学生による評価と教員による評価の間に高い相関がないことは，学生の評価力は高くないと言える。つまり，下位群の評価力も高くないと言えよう。

5.2 評価項目別の群間の分析結果と考察

5.1 でどの成績群の評価力も高くないことが確認された。そこで，学生の評価力が，評価項目によって違いがあるかをクラスカル・ウォリス検定を用いて調べた。次に，プレゼンテーション力の成績群間に差があるかを調べた。

表 6 にクラスカル・ウォリス検定の結果を示す。

表 6

クラスカル・ウォリス検定の結果

	カイ 2 乗	自由度	*p*
準備	1.034	2	.609
リサーチ	.657	2	.724
オリジナリティ	.064	2	.968
発表	2.904	2	.227
合計	.933	2	.638

表 6 から，プレゼンテーション力の上位群，中位群，下位群の間に有意差がないことがわかったが，有意差がなくても，多重比較をすると有意差が出る場合があり，また，2 群の相関係数に差があるかを調べるために，マン・ホイットニーの *U* 検定によって多重比較を行った。さらに効果量を算出した。

5.2.1 合計

表 7，表 8 に評価の「合計」の記述統計量と検定結果を示す。

表 7

評価の「合計」の記述統計量

	n	*M*	*SD*	最大値	最小値
上位群	13	.529	.184	.788	.212
中位群	13	.580	.186	.866	.313
下位群	16	.609	.198	.934	.18

表 8

評価の「合計」のマン・ホイットニーの *U* 検定結果・効果量

成績群	*U*	*W*	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
上位群 vs 中位群	74	165	-.539	.614	-.106	小	中位群＞上位群
中位群 vs 下位群	96	187	-.351	.746	-.066	無	
上位群 vs 下位群	80	171	-1.053	.308	-.207	小	下位群＞上位群

表 7 から，評価の「合計」における教員による評価と学生による評価の相関係数は，下位群が最も高く，次いで中位群，上位群が最も低いことがわかった。

表 7 と表 8 から，評価の「合計」における教員による評価と学生による評価の相関係数は，各群間に有意差はないが，上位群と中位群には効果量小 (r = -.10) で差が見られ，また，上位群と下位群の間にも効果量小 (r = -.20) で差が見られた。中位群と下位群との間には効果量もなかったことから，「合計」においては，上位群の評価力が一番低いと言える。つまり評価項目の合計とは，プレゼンテーション力を意味することから，プレゼンテーション力が一番高いとされる上位群が，プレゼンテーションの良し悪しが一番わかっていなかったことがわかった。

5.2.2 準備

表 9，表 10 に評価項目「準備」の記述統計量と検定結果を示す。

表 9

評価項目「準備」の記述統計量

	n	*M*	*SD*	最大値	最小値
上位群	13	.370	.333	.78	-.151
中位群	13	.472	.071	.555	.305
下位群	16	.496	.224	.894	.023

表 10

評価項目「準備」のマン・ホイットニーの *U* 検定結果・効果量

成績群	*U*	*W*	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
上位群 vs 中位群	77	168	-.385	.724	-.076	無	
中位群 vs 下位群	87	178	-.745	.475	-.139	小	下位群＞中位群
上位群 vs 下位群	86	177	-.789	.449	-.155	小	下位群＞上位群

表 9 から，評価項目「準備」における教員による評価と学生による評価の相関係数は，下位群が最も高く，次いで中位群，上位群が最も低いことがわかった。

表 9 と表 10 が示すように，評価項目の「準備」における教員による評価と学生による評価の相関係数は，各群間に有意差はないが，中位群と下位群の間に効果量小 (r = -.13) で差が見られ，上位群と下位群の間にも効果量小 (r = -.15) で差が見られた。上位群と中位群の間には効果量もなかったことから，「準備」の項目においては，下位群の評価力が一番高いことがわかった。

5.2.3 リサーチ

表 11，表 12 に評価項目「リサーチ」の記述統計量と検定結果を示す。

表 11

評価項目「リサーチ」の記述統計量

	n	*M*	*SD*	最大値	最小値
上位群	13	.338	.183	.622	.047
中位群	13	.206	.399	.761	-.562
下位群	16	.315	.243	.75	-.137

表 12

評価項目「リサーチ」のマン・ホイットニーの *U* 検定結果・効果量

成績群	*U*	*W*	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
上位群 vs 中位群	68	159	-.846	.418	-.166	小	上位群＞中位群
中位群 vs 下位群	86	177	-.789	.449	-.147	小	下位群＞中位群
上位群 vs 下位群	98	234	-.263	.812	-.052	無	

表 11 から，評価項目「リサーチ」における教員による評価と学生による評価の相関係数は，上位群が最も高く，次いで下位群，中位群が最も低いことがわかった。

表 11 と表 12 が示すように，評価項目の「リサーチ」における教員による評価と学生による評価の相関係数は，各群間に有意差はないが，上位群と中位群に効果量小 (r = -.16) で差が見られ，中位群と下位群の間にも効果量小 (r = -.14) で差が見られた。しかし，上位群と下位群との間には効果量もなかった。つまり，「リサーチ」の項目においては，中位群の評価力が一番低いことがわかった。

5.2.4 オリジナリティ

表 13，表 14 に評価項目「オリジナリティ」の記述統計量と検定結果を示す。

表 13

評価項目「オリジナリティ」の記述統計量

	n	*M*	*SD*	最大値	最小値
上位群	13	.326	.192	.591	-.073
中位群	13	.333	.156	.553	-.051
下位群	16	.297	.285	.774	-.352

表 14

評価項目「オリジナリティ」のマン・ホイットニーの *U* 検定結果・効果量

成績群	*U*	*W*	*Z*	*p*	効果量 *r*	効果の大きさ
上位群 vs 中位群	83	174	-.077	.960	-.016	無
中位群 vs 下位群	99	235	-.219	.846	-.041	無
上位群 vs 下位群	99	235	-.219	.846	-.044	無

表 13 から，評価項目「オリジナリティ」における教員による評価と学生による評価の相関係数は，中位群が最も高く，次いで上位群，下位群が最も低いことがわかった。

表 13 と表 14 から，評価項目の「オリジナリティ」における教員による評価と学生による評価の相関係数は，群間に差がないことがわかる。つまり，オリジナリティに対する評価力において群間における差はない。

5.2.5 発表

表 15，表 16 に評価項目「発表」の記述統計量と検定結果を示す。

表 15

評価項目「発表」の記述統計量

	n	*M*	*SD*	最大値	最小値
上位群	13	.469	.166	.792	.200
中位群	13	.587	.227	.847	-.030
下位群	16	.497	.183	.832	.220

表 16

評価項目「発表」のマン・ホイットニーの *U* 検定結果・効果量

成績群	*U*	*W*	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
上位群 vs 中位群	48	139	-1.872	.064	-.368	中	中位群＞上位群
中位群 vs 下位群	68	204	-1.579	.121	-.294	小	中位群＞下位群
上位群 vs 下位群	97	188	-.307	.779	-.061	無	

表 15 から，評価項目「発表」における教員による評価と学生による評価の相関係数は，中位群が最も高く，次いで下位群，上位群が最も低いことがわかった。

表 15 と表 16 から，評価項目の「発表」における教員による評価と学生による評価の相関係数は，各群間に有意差はないが，上位群と中位群の間に効果量中 (r = -.36) で差が見られ，また，中位群と下位群との間にも効果量小 (r = -.29) で差が見られた。しかし，上位群と下位群との間には効果量もなかった。つまり，「発表」の項目に対しては，中位群の評価力が一番高いことがわかった。

評価項目別の群間の分析結果を表 17 にまとめた。

表 17

評価項目別の群間の分析結果のまとめ

成績群	準備	リサーチ	オリジナリティ	発表	合計
上位群 vs 中位群		上位群		中位群	中位群
中位群 vs 下位群	下位群	下位群		中位群	
上位群 vs 下位群	下位群				下位群

学生の評価力において，当初はプレゼンテーション力の上位群，中位群，下位群の順で，評価力も同じ順になると予想していたが，表 17 が示すように，そのようにはならないことが確認された。本来ならば，プレゼンテーション力が高いということは，プレゼンテーションの良し悪しがわかっていると考えられることから，上位群の評価力が一番高くなることが予想された。しかし，予想とは反対に，プレゼンテーション力が低い中位群と下位群の方が，学生のプレゼンテーション力を評価する力が上位群よりあるという結果になった。そこで，なぜこのような結果になったのかについて，各群の評価傾向を調べることにより，背景となる原因を探ることを試みた。

6. 分析 2：何が学生の評価傾向に影響を及ぼすか

6.1 分析方法

各群における学生の評価傾向を調べるために，筆者が各学生のプレゼンテーションを採点して得られた得点を偏差値化し，偏差値 55 以上を高評価者，偏差値 45 から 54 を中評価者，偏差値 45 未満を低評価者とした。学生が行ったプレゼンテーションに対する学生による相互評価と，教員による相互評価との相関係数を評価群ごとに算出した。検定方法等については，これまでと同様とする。

そして，さらに詳しく学生の評価傾向を調べるため，各群の評価者群別に，教員による評価の点数と学生による相互評価の点数の差を調べた。

6.2 分析結果と考察

プレゼンテーション力の上位群，中位群，下位群それぞれの評価傾向に差があるかについてクラスカル・ウォリス検定を用いて調べた。

表 18 にクラスカル・ウォリス検定の結果を示す。

表 18
クラスカル・ウォリス検定の結果

	カイ 2 乗	自由度	*p*
上位群	.903	2	.636
中位群	2.665	2	.263
下位群	3.593	2	.165

表 18 によれば，プレゼンテーション力の各成績群における評価者群（高評価者，中評価者，低評価者）の間に有意差がないが，有意差がなくても，多重比較をすると有意差が出る場合があり，また，2 群の相関係数に差があるかを調べるために，マン・ホイットニーの *U* 検定によって多重比較を行った。さらに効果量を算出した。

6.2.1 上位群の評価傾向

表 19 に上位群の評価者別に見た記述統計量・検定結果・効果量を示す。

上位群における高評価者（偏差値 55 以上）とは評価点 40 点満点中 32 点以上をつけている学生のことであり，中評価者（偏差値 45 から 54）とは 28〜30 点をつけている学生のことであり，低評価者（偏差値 45 未満）とは 26 点以下をつけている学生のことを指す。

表 19

上位群の記述統計量・検定結果・効果量

上位群	*n*	*M*	*SD*	*U*	*W*	*Z*	*p*	効果量 *r*	効果の大きさ
高評価者	10	.226	.621	39	94	-.832	.436	-.187	小
中評価者	10	.378	.568						中評価者
中評価者	10	.378	.568	48	103	-.151	.912	-.034	無
低評価者	10	.382	.723						
高評価者	10	.226	.621	39.5	94.5	-.794	.436	-.178	小
低評価者	10	.382	.723						低評価者

表 19 から, 上位群においては, 学生によるプレゼンテーションに対する高評価者群、中評価者群、低評価者群による相互評価と教員による評価の相関は弱いことがわかった。5.1 の分析で示したように，上位群の評価力が高くないことがここでも言える。3 つの評価者間に差があるかを調べると，各評価者群間に有意差はないが，高評価者と中評価者の間には効果量小 (r = -.18) で差があり，また，高評価者と低評価者の間には効果量小 (r = -.17) で差が見られた。しかし，中評価者と低評価者の間には効果量もなかった。このことから，上位群において，高評価者より中評価者，小評価者の評価力の方が高いことがわかった。

これを教員による評価と各評価者群との実際の評価点の差を調べることによって分析した。上位群の結果を図 4 に示す。

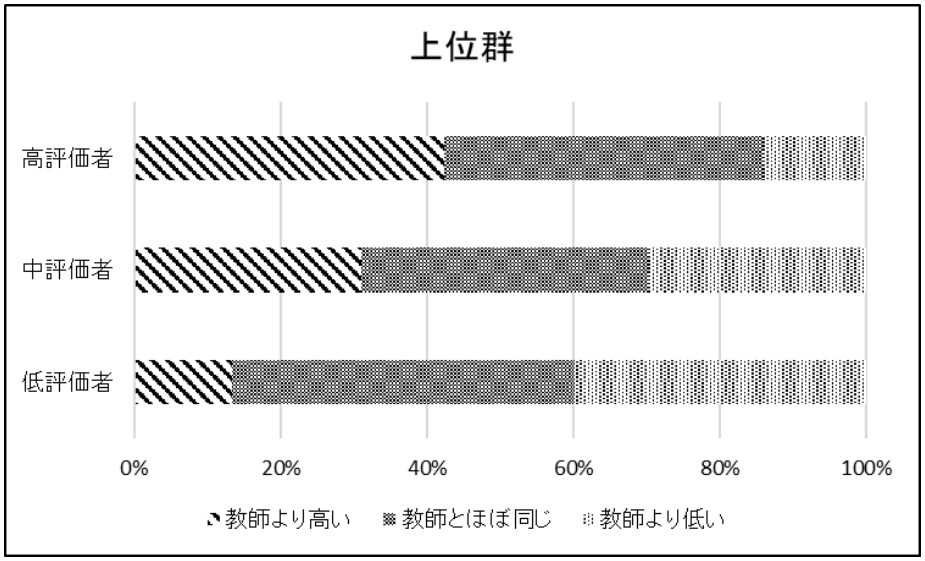


図 4. 学生による相互評価と教員による評価の評価点の差（上位群）

図 4 から，上位群の場合，高評価をつけた学生（延べ 36 名[8]）は，ほかの評価者群（中評価者＝延べ 22 名，低評価者＝延べ 8 名）と比べると，教員より高い評価点をつけるのが一番多いことがわかる。これは，上位群の学生は，自分よりプレゼンテーション力が低い学生を評価することになるため，自分のプレゼンテーション力は高いとある程度認識していると考えられ，自分と同じか，あるいは自分よりプレゼンテーションが上手いと感じた発表には，かなり高い点数をつけ，結果として甘めの評価をしていると考えられる。

反対に，教員の点数より低い点数をつける学生は，高評価者が延べ 12 名，中評価者が延べ 21 名，そして低評価者が延べ 24 名で，低評価者が一番多かった。これは，上位群の学生は，自身のプレゼンテーションと比べると，あまり出来の良い発表ではないと感じることから，点数を低くつけたと考えられる。

このように上位群の評価者には極端な評価をする傾向があると言えよう。そのため，7.1 の分析で示したように，教師の評価との相関が低くなった可能性があると考えられる。同様の結果は、De Grez（2010）にも見られる。

6.2.2　中位群の評価傾向

表 20 に中位群の評価者別に見た記述統計量・検定結果・効果量を示す。

中位群における高評価者（偏差値 55 以上）とは評価点 40 点満点中 36 点以上をつけている学生のことであり，中評価者（偏差値 45 から 54）とは 30〜34 点をつけている学生のことであり，低評価者（偏差値 45 未満）とは 28 点以下をつけている学生のことを指す。

表 20

中位群の記述統計量・検定結果・効果量

中位群	*n*	*M*	*SD*	*U*	*W*	*Z*	*p*	効果量 *r*	効果の大きさ
高評価者	5	.140	.663	16	31	-1.476	.160	-.358	中
中評価者	12	.545	.343						中評価者
中評価者	12	.545	.343	28.5	56.5	-1.142	.261	-.262	小
低評価者	7	.114	.772						中評価者
高評価者	5	.140	.663	15.5	30.5	-.326	.755	-.095	無
低評価者	7	.114	.772						

表 20 から，中位群においては，学生によるプレゼンテーションに対する高評価者群、

低評価者群による相互評価と教員による評価の相関がないことがわかった。しかし，中評価者と教員による評価の間には中程度の相関が得られた。また，評価者群間の差を調べると，各評価者群間に有意差はなかったが，高評価者と中評価者との間に効果量中 (r = -.35) で差があり，また，中評価者と低評価者との間に効果量小 (r = -.26) で差が見られた。しかし，高評価者と低評価者との間には効果量もなかった。このことから，中評価者の評価力が一番高いと言える。

図 5 に中位群における学生による相互評価と教員による評価点の差を示す。

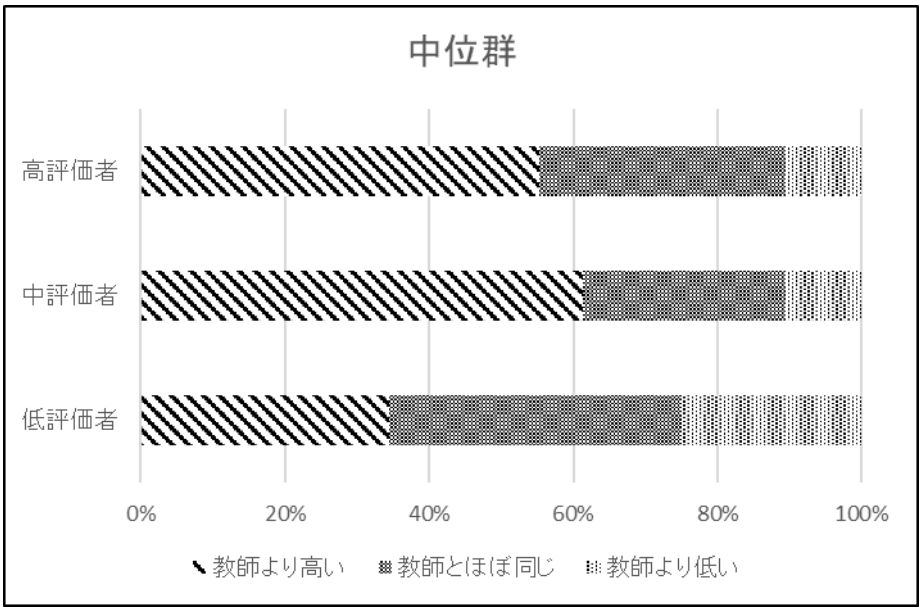


図 5. 学生による相互評価と教員による評価の評価点の差（中位群）

図 5 から，中位群の場合，教員とほぼ同じ点数をつけているのは，中評価者（延べ 26 名）が，高評価者（延べ 16 名），低評価者（延べ 13 名）と比べると，一番多いことがわかった。このことからも，中評価者の評価力が一番高いと言える。

中位群の学生は，評価の際，自分より上手なパフォーマンスを行う上位群の発表，また自分よりあまり上手くないパフォーマンスを行う下位群の発表の評価を行うことになる。そのため，それぞれ上位群の学生に対しては高い評価を，下位群の学生に対しては低い評価をしたものと考えられる。

6.2.3　下位群の評価傾向

表 21 に下位群の評価者別に見た記述統計量・検定結果・効果量を示す。

下位群における高評価者（偏差値 55 以上）とは評価点 40 点満点中 36 点以上をつけている学生のことであり，中評価者（偏差値 45 から 54）とは 30〜34 点をつけている学生のことであり，低評価者（偏差値 45 未満）とは 28 点以下をつけている学生のこ

とを指す。

表 21

下位群の記述統計量・検定結果・効果量

下位群	*n*	*M*	*SD*	*U*	*W*	*Z*	*p*	効果量 *r*	効果の大きさ
高評価者	10	.327	.430	75	211	-.264	.816	-.052	無
中評価者	16	.241	.593						
中評価者	16	.241	.593	38.5	174.5	-1.898	.057	-.38	中
低評価者	9	.604	.492						低評価者
高評価者	10	.327	.430	29	84	-1.309	.211	-.301	中
低評価者	9	.604	.492						低評価者

表 21 が示すように，下位群においては，高評価者と中評価者による相互評価と教員による評価との間には弱い相関が得られたが，低評価者との間には中程度の相関が得られた。各評価者群の間の差を見てみると，各評価者群間に有意差はなかったが，中評価者と低評価者の間には効果量中 (r = -.38)で差が見られ，また高評価者と低評価者との間にも効果量中 (r = -.30) で差が見られた。しかし，高評価者と中評価者との間には効果量もなかった。このことから，下位群においては，低評価者の評価力が一番高いと言える。

図 6 に下位群における学生による相互評価と教員による評価点の差を示す。

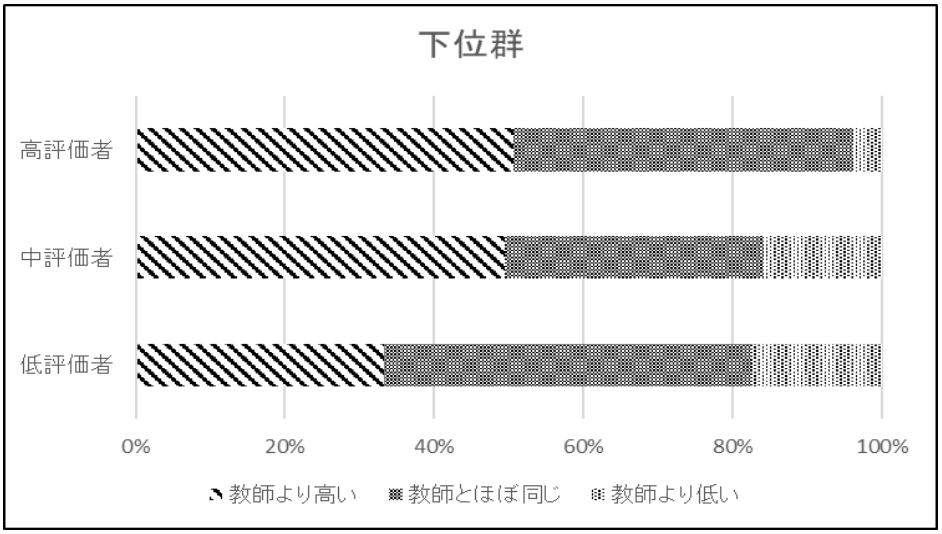


図 6. 学生による相互評価と教員による評価の評価点の差（下位群）

図 6 から，下位群の低評価者は，教員がつける点数とほぼ同じ点数をつける学生は延べ 28 名で，教員より高い点数をつける学生（延べ 19 名），教員より低い点数をつける学生（延べ 10 名）と比べると一番多かった。このことからも，下位群の低評価者が一番評価力が高いと言える。下位群の低評価者といえば，プレゼンテーション力が自分と同じくらいか，あるいは自分より低いと思った学生に対する評価であるが，下位群の学生自体，クラスの中ではプレゼンテーション力が低い学生になるので，自分よりプレゼンテーションがうまい学生を見ることで，自分のプレゼンテーション力のよくない点に気づき，そのパフォーマンスと同じあるいは似たようなプレゼンテーションを見たときには，何がよくないプレゼンテーションなのかということがよくわかったことで，甘すぎず，また厳しすぎない評価ができたと考えられる。

また，下位群においては，上位群や中位群と比べると，高めの点数をつける学生が一番多く見受けられた（上位群高評価者延べ 36 名，中位群高評価者延べ 26 名，下位群高評価者延べ 39 名）。下位群の学生が評価するのは，自分より出来の良いプレゼンテーションということになる。このことが，プレゼンテーションを上手いと思わせ，あるいは逆に劣等感を感じさせ，高めの評価をしたと考えられる。

上記の結果から，プレゼンテーション力のどの成績群においても，高評価者は，教員より高い点数をつける傾向があることがわかった。つまり，高い点数をつけるということは，その学生の発表は上手いと判断したわけであるから，学生全体の傾向として，学生は自分より上手い発表だと感じたときは，かなり甘めの評価をすることが考えられる。

7. 全体の考察

5.1 では, 学生のプレゼンテーション力の高さが評価力に影響を及ぼすかを確かめるために，学生による相互評価と教員による評価の相関係数を求め分析した結果，すべての成績群の評価の合計点において中程度の相関が得られたことから，どの成績群の学生も評価力が低いことがわかった。つまり，学生のプレゼンテーション力が高いことが，必ずしも評価力が高いということを証明できなかった。

学生による相互評価と教員による評価の相関係数を評価項目別に見てみると,「リサーチ」と「オリジナリティ」に比べて、「準備」と「発表」では相関が高くなる傾向が見られた。このことは，笠巻（2016）の結果とも合致する。これは,「準備」に関しては, しっかりと準備をして臨んでいる発表と十分な準備をせずに臨んだ発表の違いは, 学生にとってわかりやすいものであったと思われる。また,「発表」の仕方についても, 発音，声の大きさ，速度，アイコンタクト，明瞭さなどは，その良し悪しが見た目にはっきりとわかるものであることから，学生にとっては評価しやすい項目であったと考えられる。一方，リサーチとオリジナリティという評価項目は，いわゆる内容に関

わる項目である。学生の発表内容の良し悪しの判断には，評価者の予備知識，つまりその分野についてよく知っているかが影響していると思われる。具体的には，一般の人が素晴らしいと思う発表であっても，その分野のことをよく知っている評価者が聴けば，評価が低くなる場合がある。その逆も同じで，その分野のことをよく知らない評価者が聴けば，評価を甘く，また高めにしてしまう場合がある。つまり，学生にとって，クラスメイトの一人ひとり違う発表の内容を評価することは難しいと思われる。このため，教員による評価との相関が低くなったと考えられる。つまり，4 つの評価項目の中でも，特にプレゼンテーションの内容に関わる項目の評価力は低いと言えよう。

また，5.2 では，学生の評価力が成績群間で差があるかについて調べたが，学生の評価力は，プレゼンテーション力の上・中・下と同じ順にはならないことが確認された。そして，成績群の中で，一番プレゼンテーション力が高いとされる上位群の評価力が一番低いことが判明した。本来，プレゼンテーション力が高いということは，プレゼンテーションの良し悪しがよくわかっていると予想されることから，上位群の評価力が一番高くなることが予想されたが，予想とは反対に，学生のプレゼンテーション力を評価する力は，上位群が一番低い結果となった。すなわち，学生のプレゼンテーション力は学生の評価力とは関係がない可能性があると言えるだろう。

分析 2 は，分析 1 で，学生の評価力が高くないこと，そして学生のプレゼンテーション力が学生の評価力とは関係がない可能性があるという結果から，学生の評価傾向を詳しく調べたものである。各群の評価傾向を調べると，学生の評価傾向として，学生全体としては，教員の評価と比べると甘めの評価を行う傾向があることがわかった。このことは，Cheng & Warren（2005），Fukazawa（2010），笠巻（2016）の結果とも合致する。さらに成績群別に傾向を調べると，上位群の学生は厳しめに評価を行う傾向があり，逆に下位群の学生は高めの点数をつける傾向があることがわかった。このことは，学生は，自分の発表の出来を元に，その出来とクラスメイトの発表の出来を比べ，評価してしまう傾向があると考えられる。つまり，学生の評価はかなり主観的なものになってしまい，そのため，評価の信頼性は低い（Freeman, 1995; 笠巻, 2016）と言えよう。

このような結果となった原因として，次のような点が考えられる。まず一つ目として，現行の評価シートには，評価項目と判定基準が記されてはいるものの，評価の観点が明確には書かれていない点が挙げられる。また，教師も評価シートの説明のみ行っており，判定基準や評価の観点において詳しく説明していなかったことから，学生間でバラつきが生じ，主観的な評価をしていたことが原因の一つと考えられる。学生により客観的な評価をさせるためには，評価項目と判定基準，評価の観点を明確に記したルーブリックを作成する必要があると考える。

もう一つの原因として，学生の評価経験の浅さが考えられる。評価経験の有無が評価の信頼性や一貫性に大きく影響するということは，Weigle（1994），山西（2004）の研究でもすでに述べられている。評価を行えば行うほど，評価がより正確になると言われている（Nakamura, 2002）。今回，学生の相互評価の経験の有無は調べてはいないが，本研究を行ったのは，1 年生の前期における中間発表での評価であることから，相互評価の経験はあまりないと考えられる。評価の経験が少ないことから，学生の評価力は非常に未熟と言わざるを得ない。

この他に考えられる原因として，発表内容を理解するのに十分な英語力がない学生が多かったために，相関係数が中程度になったと思われる。また，発表内容に詳しい学生とそうでない学生が混在していたことも，教員による評価と学生による相互評価の相関が中程度になった一因であろう。学生の英語力が学生の評価力に影響を及ぼすかについては第4章で,評価者の予備知識の有無が学生の評価力に影響を及ぼすかについては第 5 章で検証する。

その他の原因として，中間発表時において，教員と違い，学生は評価のみに徹することはできない。学生の中には，同じ日に自分の発表を控えている者もいるからである。中にはクラスメイトの発表の評価どころではないと感じている学生もいるかもしれない。そのような学生による評価の信頼性は高いとは言い難い。これについては，学生による相互評価の実施方法も検討する必要があるであろう。

最後に，先行研究と本研究の結果を踏まえて，スピーチ・プレゼンテーションのクラスにおける学生による相互評価の活用の意義について述べる。

相互評価を行うことにより，クラスメイトによる良い発表もそうでない発表もただ受け身的に聴くのではなく，プレゼンテーションの良し悪しを批判的に聴くことになる。まだ評価経験が浅く，そのため評価力が低い学生であっても，適切なトレーニングを行い，繰り返し相互評価を行うことにより，学生の批判的思考が高まり，プレゼンテーションの良し悪しがわかるようになる。すなわち，学生に評価力がついてくるのである。そして，将来自律した学習者になるために必要な判断力を身に着けさせる目的で，相互評価が活用されることにより，学生を自律した学習者に育てることができると筆者は考えている。さらには，学生による相互評価と教員による評価の信頼性を算出し，また学生の評価傾向を調べることは，主観的になりかねない教員による評価を見直す良い機会ともなり得る。つまりは，学生の評価力をあげるために必要となるだけではなく，教員の評価力をあげることにもつながると言えよう。

また，すでに報告されている相互評価における様々な利点（Brown, 1998; Goh & Burns, 2012; Fukazawa, 2010; Luoma, 2004 ; Nakamura, 2002; 三木・笠巻, 2017）から，授業内活動の一つとして用いられることには筆者も同意見である。しかし，評価にかかる時間や労力といった教師の負担を軽減するとの先行研究（Brown, 1998; Fukazawa,

2010; Luoma, 2004; Okuda & Otsu, 2010）があるが，筆者はそうは考えない。なぜなら，学生の相互評価を集計し，教員の評価に合計する作業は，時間や労力を勘案しても，決して楽なものではないからである。むしろ，負担は大きいと考える。

学生による相互評価の良し悪しを，ただ単に教員による評価との一致の度合いだけで，判断する訳ではない。むしろ，学生による相互評価が持つ他の分野における肯定的な面によく考慮してなされるべきであるという Cheng & Warren（2005）の指摘に筆者は同意する。他の分野とは，学生の相互評価が学習目標を明確にしたり，学生のモチベーションを高めたりする効果があるということであろう。

注

1. 「プレゼンテーション力」とは，学生一人ひとりの発表に対する，準備，リサーチ，オリジナリティ，発表の 4 つの評価項目から成る，教師による評価点（4 つの項目の合計点）を意味する。以後，単にプレゼンテーション力と記す。
2. テキストは，次の教科書が使用された。鈴木佑治. (2008). 『Do Your Own Project in English Volume 1: プロジェクト発信型英語 Vol.1 』郁文堂　但し，2017 年度からはこの教科書は参考書としている。
3. 参加者には，研究の目的，内容についての説明を行い，本授業から得られたデータを使用し，結果を公表することに対する同意を得た。
4. 自己評価をつけていない学生と，すべての評価項目に対して同じ点数をつけている学生の相関係数は算出されなかった。
5. 学生一人一人の点数は，授業担当者には知らされていない。
6. 10 段階評価は「プロジェクト発信型プログラム」では行っておらず，本研究における筆者担当クラスのみにて実施した。
7. 「プロジェクト発信型プログラム」では，学生の発表の評価は基本担当教員一人で行うことになっている。本研究は共同研究者を得ることができず，筆者一人で行わざるを得なかった。しかし，同プログラムでは，複数の教員がそれぞれ同内容のクラスを担当しているため，今後は筆者の担当クラス以外にも対象を拡げ，複数の教員で評価を行う必要性を感じている。
8. 評価をつけた学生の延べ人数のことであり，実際の学生の人数とは異なる。

第 4 章　研究 2 : 学生の「英語力」が評価者としての学生の評価力に影響を及ぼすか？

1. はじめに

学生の評価力に影響を及ぼすとされる４つの要因（プレゼンテーション力, 英語力, 予備知識，評価者トレーニング）の一つである，学生のプレゼンテーション力が評価者としての学生の評価力に影響を及ぼすかを検証した第 3 章（研究 1）では，各成績群の評価傾向は知りうることができたが，なぜプレゼンテーション力の高い学生の評価と教師の評価の相関が一番低くなったかについて，明確な原因を突き止めることはできなかった。

研究 2 では，上述の要因の一つである，学生の英語力の高低によって，プレゼンテーションに対する学生による相互評価と教員による評価との相関が変わるかを調べ，英語力が学生の評価力に影響を及ぼすかを検証する。

2. 英語力が評価に及ぼす影響に関する先行研究

Shimura (2006)は，英語力の違う３つのグループ（上位群：TOEFL550 点以上，中位群：TOEFL480～550，下位群：英検３級から準２級）に対して，学生の英語力が学生による評価に及ぼす影響を調べた。その結果，英語力の中位群の学生の評価が，教師による評価に一番近く，続いて下位群，次に上位群の順で，教員による評価とは一致しにくくなることがわかった。そして，必ずしも，英語力の高い学生による評価と教師による評価との相関が一番高くなるわけではないと報告している。

また，学生の評価傾向として，中位群の学生の評価が最も厳しく，下位群の学生が最も甘い評価を行う傾向があり，その理由として，下位群の学生は英語力がないために，発表内容がよくわからないためとしている。また，上位群の学生は，評価点の幅が小さく，誰に対しても高い点数をつける傾向があり，その理由として，自分の英語力がある程度高いことは認識しているが，同じクラスの学生の英語力も高いと思っているためとしている（Shimura, 2006）。

さらに，下位群の学生は，スライドなどの visual aids のような，見た目に良し悪しがわかりやすい項目で評価を行う傾向があるため，英語によるプレゼンテーションの内容や構成を評価させることは難しいとしている。また，上位群の学生は，５段階評価で評価をつけることを嫌い，むしろコメントを書いて評価する方をより好む傾向があることから，評価シートを一律ではなく，英語力に応じて教員がつくるべきだと主張している（Shimura, 2006）。

3. 研究 2 の目的

研究 2 では，評価者としての学生の評価力に影響を及ぼすと考えられる４つの要因のうち，学生の英語力が評価力に影響を及ぼすかを調べることと，英語力別にみた学生の評価傾向を調べることを目的とする。

4. 研究 2 における指導と評価方法

4.1 指導

指導内容と授業の流れについては，p. 10, 11 で述べたので，ここでは省略する。

4.2 評価方法

評価方法については，p. 11, 12 で述べたので，ここでは省略する。

5. 研究方法

5.1 参加者

研究 2 の参加者は、滋賀県内の私立大学における筆者担当クラスの１年生４クラスの計 61 名の内, 学生一人ひとりのパフォーマンスに対する教員による評価と学生による評価の相関係数が何らかの事情で欠けている学生と，相関係数が外れ値となった学生 19 名を除いて，42 名を分析対象とした。参加者の英語力は，TOEIC IP テスト 290 点～680 点で，平均は 464 点であった。

5.2 手続き

手続きについては，p. 12 で述べたので，ここでは省略する。

5.3 分析 1: 学生の英語力は評価者としての学生の評価力に影響を及ぼすか。

教員は優れた発表には高い評価をつけ，あまり良くない発表には低い評価をしているが，学生が正しく評価できるかどうかを調べるために，学生もまた教員による評価と同じように，優れた発表には高い評価を，あまりよくない発表には低い評価をしているかを調べた。

5.3.1 分析方法

各学生の英語力を、TOEIC IP スコアに基づいて偏差値化し、偏差値 55 以上（TOEIC IP 510～680 点）を上位群，偏差値 45 から 55 未満（TOEIC IP 430～505 点）を中位群，偏差値 45 未満（TOEIC IP 290～420 点）を下位群とし，学生が行ったプレゼンテーションに対する学生による相互評価と，教員による相互評価との相関係数を成績群ごとに算出した。学生一人ひとりのパフォーマンスに対する学生による評価の合計点の正規性の検定を行ったところ，（表 1）が示す結果になった。次に，英語力の上位群，中位群，下位群の間に差があるかについてクラスカル・ウォリス検定を用いて調べ，その後，多重比較を行なった。2 群の相関係数に差があるかを調べるために，マン・ホイットニーの *U* 検定を行い，効果量を算出した。

表 1

正規性の検定結果

	Kolmogorov-Smirnov(a)		
	統計量	自由度	有意確率
上位群	.156	180	.000
中位群	.146	180	.000
下位群	.144	180	.000

6. 結果と考察

6.1 群別の分析結果と考察

6.1.1 上位群

表 2 に，英語力の上位群における，学生による相互評価と教員による評価との相関係数の記述統計量を示す。

表 2

上位群の記述統計量 (*n*=18)

	M	*SD*	最小値	最大値
準備	.408	.287	-.151	.894
リサーチ	.352	.211	.066	.761
オリジナリティ	.265	.260	-.376	.608
発表	.556	.190	.200	.847
合計	.575	.195	.212	.889

英語力上位群においては、教師による評価と学生による評価との相関は、評価の合計点において中程度の相関(r = .575)が得られた。しかし，評価項目別に見てみると、発表(r = .556)と準備(r = .408)においては中程度の相関が得られたが、リサーチ(r = .352)、オリジナリティ(r = .265)においては弱い相関が得られた。

また, 評価の合計（SD = .195）においては, 強い相関（r = .889）から弱い相関（r = .212）になるなど，相関係数が 3 段階に大きくばらついていることがわかった。また，評価項目別に見ると，リサーチ（SD = .211）においては，強い相関（r = .761）から相関がない関係（r = .066）に, オリジナリティ（SD = .260）においては, 中程度の相関（r = .608）からマイナスの関係（r = −.376），発表（SD = .190）においては，強い相関（r = .847）から弱い相関（r = .200）になるなど，相関係数が 3 段階に大きくばらついていることがわかった。さらに，準備（SD = .287）においては，強い相関（r = .894）からマイナスの関係（r = −.151）になり，相関係数が 4 段階に大きくばらついていることがわかった。

これらの結果から，英語力上位群の学生の評価と教員による評価の相関は高くないことがわかった。学生による評価と教員による評価の間に高い相関がないことは，学生の評価力は高くないと言える。つまり，英語力上位群の評価力は高くないと言えよう。

6.1.2 中位群

表 3 に，英語力の中位群における，学生による相互評価と教員による評価との相関係数の記述統計量を示す。

表 3

中位群の記述統計量 (n=11)

	M	SD	最小値	最大値
準備	.493	.208	-.053	.705
リサーチ	.385	.206	.058	.697
オリジナリティ	.371	.129	.157	.533
発表	.458	.234	-.015	.816
合計	.589	.166	.290	.794

英語力中位群においては、教師による評価と学生による評価との相関は、評価の合計点において中程度の相関(r = .589)が得られた。しかし，評価項目別に見てみると、準備(r = .493)と発表(r = .458)においては中程度の相関が得られたが、リサーチ(r = .385)とオリジナリティ(r = .371)においては弱い相関が得られた。

また, 評価の合計（*SD* = .166）においては, 強い相関（*r* = .794）から弱い相関（*r* = .290）になるなど，相関係数が 3 段階に大きくばらついていることがわかった。また，評価項目別に見ると，オリジナリティ（*SD* = .129）においては，中程度の相関（*r* = .533）から相関がない関係（*r* = −.157）, になり, 相関係数が 3 段階に大きくばらついており，リサーチ（*SD* = .206）においては，強い相関（*r* = .697）から相関がない関係（*r* = .058）になるなど，相関係数が 4 段階に大きくばらついていることがわかった。さらに，準備（*SD* = .208）においては，強い相関（*r* = .705）からマイナスの関係（*r* = −.053）になり, 発表（*SD* = .234）においては, 強い相関（*r* = .816）からマイナスの関係（*r* = −.015）になるなど，相関係数が 5 段階に大きくばらついていることがわかった。

これらの結果から，英語力中位群の学生の評価と教員による評価の相関は高くないことがわかった。学生による評価と教員による評価の間に高い相関がないことは，学生の評価力は高くないと言える。つまり，英語力中位群の評価力も高くないと言えよう。中位群の学生から見れば，自分より英語力の高い学生の発表も，また逆に低い学生の発表も聴くことになる。自分より英語力の高い学生の発表を聴けば，とてもうまい発表であると感じることから，教員による評価と比べると甘めの評価を行い，反対に，自分より英語力の低い学生の発表には，あまりうまくない発表であると感じてしまうことから，教員による評価と比べると低めの評価をした可能性がる。そのため，どの項目においても，教員による評価と学生による評価の相関係数のばらつきが大きくなったと考えられる。

表 4 に，英語力の下位群における，学生による相互評価と教員による評価との相関係数の記述統計量を示す。

表 4

下位群の記述統計量 (*n*=13)

	M	*SD*	最小値	最大値
準備	.442	.179	.023	.752
リサーチ	.129	.375	-.582	.750
オリジナリティ	.352	.217	-.052	.774
発表	.493	.166	.220	.752
合計	.551	.223	.180	.934

英語力下位群においては、教師による評価と学生による評価の相関は、評価の合計点において中程度の相関(*r* = .551)が得られた。評価項目別に見てみると、発表(*r* = .493)と準備(*r* =.442)においては中程度の相関が得られたが、リサーチ(*r* = .129)、オリジナリティ(*r* = .352)においてはそれぞれ弱い相関が得られた。

また，評価の合計（*SD* = .223）においては，強い相関（*r* = .934）から相関がない関係（*r* = .180）になるなど，相関係数が 4 段階に大きくばらついていることがわかった。また，評価項目別に見ると，リサーチ（*SD* = .375）においては，強い相関（*r* = .750）からマイナスの関係（*r* = −.582）になり，オリジナリティ（*SD* = .217）においても，強い相関（*r* = .774）からマイナスの関係（*r* = −.052），になるなど，相関係数が 5 段階に大きくばらついていることがわかった。さらに，準備（*SD* = .179）においては，強い相関（*r* = .752）から相関がない関係（*r* = −.023）になり，相関係数が 4 段階に大きくばらついており，発表（*SD* = .166）においては，強い相関（*r* = .752）から弱い相関（*r* = .220）になるなど，相関係数が 3 段階に大きくばらつくなど，英語力下位群における教師による評価と学生による評価の相関は，評価の合計においても，どの評価項目においても，かなり大きくばらついていることがわかった。

これらの結果から，英語力下位群の学生の評価と教員による評価の相関は高くないことがわかった。学生による評価と教員による評価の間に高い相関がないことは，学生の評価力は高くないと言える。つまり，英語力下位群の評価力も高くないと言えよう。

6.2　評価項目別の群間の分析結果と考察

6.1 でどの成績群の評価力も高くないことが確認された。そこで，学生の評価力が，英語力の成績群間に差があるかを調べた。次に，学生の評価力が評価項目によって違いがあるかを調べた。

表 5 にクラスカル・ウォリス検定の結果を示す。

表 5

クラスカル・ウォリス検定の結果

	ｶｲ 2 乗	自由度	*p*
準備	1.353	2	.522
リサーチ	3.684	2	.165
オリジナリティ	1.076	2	.598
発表	1.587	2	.450
合計	.266	2	.885

英語力の上位群、中位群、下位群の間に差があるかどうかを、クラスカル・ウォリス検定を用いて調べたところ、有意差はなかった。クラスカル・ウォリス検定で有意差はなくても、多重比較をすると有意差が出る場合があり、また、２群の相関係数の差がどの程度の大きさであるかを調べるために、マン・ホイットニーの *U* 検定によ

る多重比較を行った。さらに効果量を算出した。

6.2.1 合計

表 6，表 7 に評価の「合計」の記述統計量と検定結果を示す。

表 6

「合計」の記述統計量

	n	*M*	*SD*	最小値	最大値
上位群	18	.575	.195	.212	.889
中位群	11	.589	.166	.29	.794
下位群	13	.551	.223	.18	.934

表 7

「合計」のマン・ホイットニーの *U* 検定結果・効果量

					効果量	効果の	
	U	*W*	*Z*	*p*	*r*	大きさ	優劣関係
上位群 vs 中位群	95.5	266.5	-.157	.877	-.03	なし	
中位群 vs 下位群	63	154	-.492	.649	-.101	小	中位群＞下位群
上位群 vs 下位群	107.5	198.5	-.380	.708	-.069	なし	

表 6 と表 7 から，評価の「合計」における教員による評価と学生による評価の相関係数は，各群間に有意差はないが，中位群と下位群との間に効果量は小（*r* = -.10）で差が見られた。しかし，上位群と中位群，上位群と下位群の間には効果量もなかった。このことから，評価の合計点においては，中位群の評価力が一番高いと言えよう。このことは，英語力上位群の学生は自分より英語力の低い学生による発表を評価することになるため，評価が厳し目になったと考えられる。一方，英語力下位群の学生は，自分より英語の上手い発表を聴くことになるため，評価が甘目になったと考えれらる。中位群の学生は自分より英語の上手い上位群の発表には高めの評価をし，逆に英語の下手な下位群の発表には低く評価をしていることから，評価力が一番高くなったと考えられる。この点については分析 2 で詳細を分析する。

6.2.2 準備

表 8，表 9 に，評価項目「準備」の記述統計量と検定結果を示す。

表 8

評価項目「準備」の記述統計量

	n	*M*	*SD*	最小値	最大値
上位群	18	.408	.287	-.151	.894
中位群	11	.493	.208	-.053	.705
下位群	13	.442	.179	.023	.752

表 9

評価項目「準備」のマン・ホイットニーの *U* 検定結果・効果量

	U	*W*	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
上位群 vs 中位群	77	248	-.989	.340	-.184	小	中位群＞上位群
中位群 vs 下位群	53	144	-1.072	.303	-.219	小	中位群＞下位群
上位群 vs 下位群	112	283	-.200	.859	-.036	なし	

表 8 と表 9 から，評価項目の「準備」における教員による評価と学生による評価の相関係数は，各群間に有意差はないが，上位群と中位群との間に効果量は小（*r* = -.18）で差が見られ，上位群より中位群の方が「準備」における評価力は高いことがわかる。また，中位群と下位群との間にも効果量は小（*r* = -.21）で差が見られ，下位群より中位群の方が評価力が高いことがわかる。しかし，上位群と下位群の間には差がなかったことから，「準備」においては，中位群の評価力が一番高いと言える。

「準備」がよくできている発表というものは，とても流れのよいスムーズな発表になる。そして，英語力の高い学生は，スムーズに発表できることから，たどたどしい英語や詰まりながら話す発表は，準備不足と考えて厳しめの評価をしたと考えられる。これが教員による評価との相関に影響を及ぼしたと考えられる。一方，下位群の学生がスムーズな発表を聴けば，とてもよく準備された発表と思えることから，高めの評価点をつけたと考えられ，その結果，教員より甘目に評価したと考えられる。これが教員による評価との相関が，中位群より低くなった原因であろう。中位群の学生は，スムーズな発表には高い評価を，そうでない発表には低い評価をつけたと考えられ，評価力が一番高くなったと考えられる。この点については分析 2 で詳細を分析する。

6.2.3 リサーチ

表 10，表 11 に，評価項目「リサーチ」の記述統計量と検定結果を示す。

表 10

評価項目「リサーチ」の記述統計量

	n	*M*	*SD*	最小値	最大値
上位群	18	.352	.211	.066	.761
中位群	11	.385	.206	.058	.697
下位群	13	.129	.375	-.582	.75

表 11

評価項目「リサーチ」のマン・ホイットニーの *U* 検定結果・効果量

	U	*W*	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
上位群 vs 中位群	87	258	-.539	.611	-.101	小	中位群＞上位群
中位群 vs 下位群	40	131	-1.825	.072	-.373	中	中位群＞下位群
上位群 vs 下位群	81	172	-1.441	.157	-.259	小	上位群＞下位群

表 10 と表 11 から，評価項目の「リサーチ」における教員による評価と学生による評価の相関係数は，各群間に有意差はないが，上位群と中位群の間に効果量は小（*r* = -.10）で差が見られ，「リサーチ」おける評価力は，上位群より中位群の方が高いことがわかる。また，上位群と下位群の間にも効果量は小（*r* = -.25）で差が見られ，下位群より上位群の方が評価力が高いことがわかる。また，中位群と下位群の間にも効果量は中（*r* = -.37）で差が見られ，下位群より中位群の方が評価力が高いことがわかる。これらの結果から，評価項目の「リサーチ」においては，中位群の評価力が一番高く，次に上位群，そして下位群が一番低いことがわかる。

英語力中位群の学生は，発表者のリサーチした内容をよく理解できたものには高い評価を，反対にあまりよく理解できなかった発表には低い評価をしたことで，教員による評価との相関が一番高くなったと考えられる。英語力上位群の学生はリサーチした内容の発表も上手くできると考えられ，その学生が自分より英語力の低い学生の発表を聴くと，理解しづらい英語であったり，リサーチした内容の発表の仕方も自分より下手に見えることから，評価が厳しめになったと考えられる。これが教員との評価の相関が中位群より低くなった原因であろう。一方，下位群の学生は，リサーチした内容をよく理解できないまま評価をしたと考えられる。一般に内容がよく理解できるものに対しては評価は厳しくなり，反対によくわからない内容の発表には評価が甘くなることから，下位群の学生の評価は甘めになったと考えられる。これが教員との評価の相関が上位群，中位群より低くなった原因と考えられる。このことは，「リサーチ」の項目は内容に関わることであることから，発表内容をよく理解する必要があり，そ

のためには十分な英語力が必要と考えられる。この点については分析 2 で詳細を分析する。

6.2.4 オリジナリティ

表 12，表 13 に評価項目「オリジナリティ」の記述統計量と検定結果を示す。

表 12
評価項目「オリジナリティ」の記述統計量

	n	*M*	*SD*	最小値	最大値
上位群	18	.265	.260	-.376	.608
中位群	11	.371	.129	.157	.533
下位群	13	.352	.217	-.052	.774

表 13
評価項目「オリジナリティ」のマン・ホイットニーの *U* 検定結果・効果量

	U	*W*	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
上位群 vs 中位群	77	248	-.989	.340	-.184	小	中位群＞上位群
中位群 vs 下位群	69	160	-.145	.910	-.03	なし	
上位群 vs 下位群	99	270	-.721	.489	-.13	小	下位群＞上位群

表 12 と表 13 から，「オリジナリティ」における教員による評価と学生による評価の相関係数は，各群間に有意差はないが，上位群と中位群の間に効果量は小（r = -.18）で差が見られ，中位群の評価力の方が高いことがわかる。また，上位群と下位群の間にも効果量は小（r = -.13）で差が見られ，上位群の評価力の方が高いことがわかる。しかし，中位群と下位群の間には差はなかったことから，「オリジナリティ」においては，上位群の評価力が一番低いことがわかった。

「オリジナリティ」という項目を正しく評価するには，発表内容をよく理解する必要がある。そのためには英語力が必要となる。英語力上位群はクラスメイトの発表内容を十分理解できるだけの英語力があると判断できることと，上述の理由から，上位群の学生による評価は厳しめの評価になったと考えられる。これが教員による評価との相関が，中位群，下位群より低くなった原因であろう。逆に，中位群，下位群の学生は発表内容を十分に理解できていないと考えられることから，評価が甘目になったと考えられる。これが教員による評価との相関に影響を及ぼしたと考えられる。この

点については分析 2 で詳細を分析する。

6.2.5 発表

表 14，表 15 に評価項目「発表」の記述統計量と検定結果を示す。

表 14

評価項目「発表」の記述統計量

	n	*M*	*SD*	最小値	最大値
上位群	18	.556	.190	.2	.847
中位群	11	.458	.234	-.015	.816
下位群	13	.493	.166	.22	.752

表 15. 評価項目「発表」のマン・ホイットニーの *U* 検定結果・効果量

	U	*W*	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
上位群 vs 中位群	73.5	139.5	-1.146	.256	-.213	小	上位群＞中位群
中位群 vs 下位群	67	133	-.261	.820	-.054	なし	
上位群 vs 下位群	94	185	-.921	.373	-.166	小	上位群＞下位群

表 14 と表 15 から，評価項目の「発表」における教員による評価と学生による評価の相関係数は，各群間に有意差はないが，上位群と中位群の間に効果量は小（*r* = -.21）で差が見られ，上位群の評価力の方が高いことがわかる。また，上位群と下位群の間にも効果量は小（*r* = -.16）で差が見られ，上位群の評価力の方が高いことがわかる。しかし，中位群と下位群の間に差はなかったことから，「発表」に対する評価力は上位群が一番高いことがわかる。

このことは，英語力の上位群というのは，英語によるプレゼンテーション力も高いと予想される。プレゼンテーション力が高いということは，発表の仕方が良し悪しをよく理解できていると考えられることから，上手い発表には高い評価を，またそうでない発表には低い評価をつけたことから，評価力が高くなったと言えよう。これが教員による評価との相関に影響を及ぼしたと考えられる。一方，中位群，下位群の学生は，自分よりプレゼンテーション力の高いつまり，上手な発表を聴くことになる。そのため，自分より上手な発表に対して，かなり甘目の評価を行ったと考えられる。これが教員による評価との相関が，上位群より低くなった原因であろう。

評価項目別の群間の結果を表 16 にまとめた。

表 16

評価項目別の群間の分析結果のまとめ

	準備	リサーチ	オリジナリティ	発表	合計
上位群 VS 中位群	中位群	中位群	中位群	中位群	
中位群 VS 下位群	中位群	中位群		中位群	中位群
上位群 VS 下位群		上位群	下位群	上位群	

英語力が高いということは、発表内容をよく理解して評価を行うため、教員による評価との相関が高くなるものと予想されたが、表 16 が示すように，そのような結果にはならなかった。これは，英語力上位群の学生は，自分より英語が下手でわかりにくい発表を聴いて評価することになることから，評価が厳しめになったと考えられる。一方，英語力下位群の学生は，自分より英語が上手な発表を聴いて評価することになるため，甘目の評価をしたと考えられる。つまり，学生は自分の英語力と比較して，クラスメイトの発表の評価を行っていると考えられ，彼らが行う評価はかなり主観的なものになっている可能性があることから，その評価の信頼性は低いと言えよう。このことから、学生の英語力の高低だけが、学生の評価力に影響を及ぼすことはなく，学生による相互評価の信頼性を高めることもできないと言えよう。

なぜこのような結果になったのか，学生の評価傾向を調べることで，その原因を探った。

7. 分析 2：何が学生の評価傾向に影響を及ぼすか

7.1 分析方法

各群における学生の評価傾向を調べるために，各群の学生の評価の合計点をヒストグラムを用いて示した。そして，さらに詳しく学生の評価傾向を調べるために，各群の評価者群別に，教員による評価の合計点の点数と学生による相互評価の合計点の点数の差を調べた。

7.2 分析結果と考察

7.2.1 教員の評価傾向

図 1 に教員の評価傾向を示す。

図 1 から，教員による評価は，高い点数も低い点数もつけていることがわかる。つまり，良いプレゼンテーションには高い点数をつけ，そうでないプレゼンテーションには低い点数をつけていると言える。

7.2.2　学生の評価傾向

図 2～4 に各評価者群の評価傾向を示す。

学生による評価は，図 2，3，4 が示すように，低めの点数をつける学生は少なく，高めの点数をつける学生が多いことがわかる。つまり，学生はやや甘目の評価をする傾向があることがわかる。

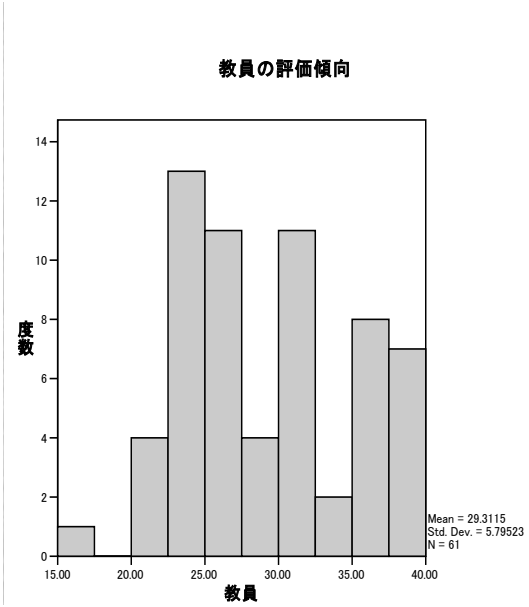


図 1.教員による評価の合計点

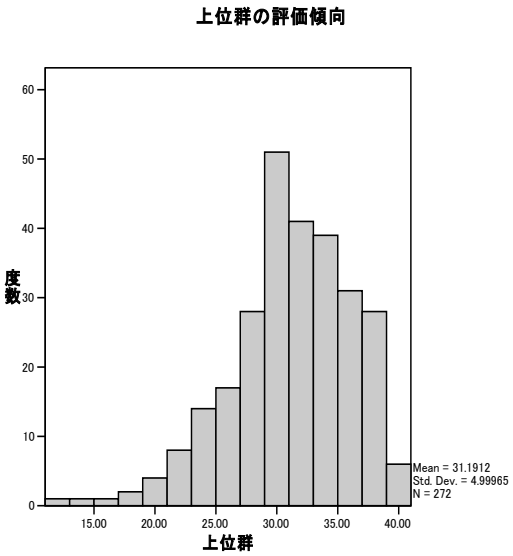


図 2.上位群よる評価の合計点

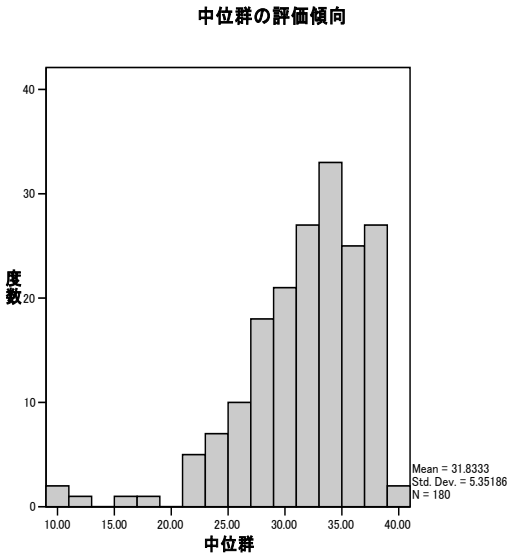


図 3.中位群による評価の合計点

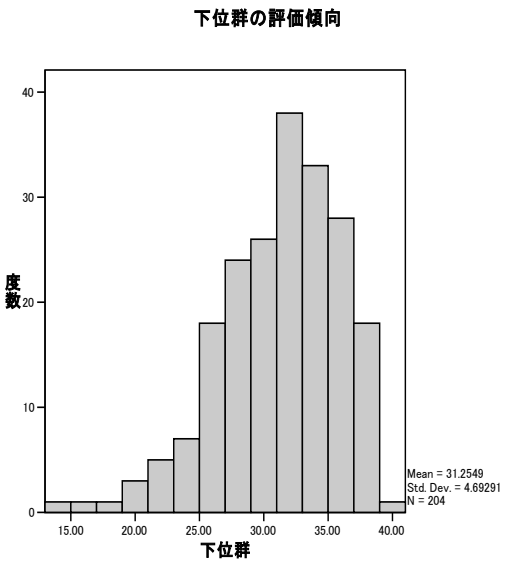


図 4.下位群による評価の合計点

次に，学生の評価傾向をさらに詳しく調べるため，教員による評価と各評価者群との実際の評価点の差を調べることによって分析した。その結果を図 5 に示す。

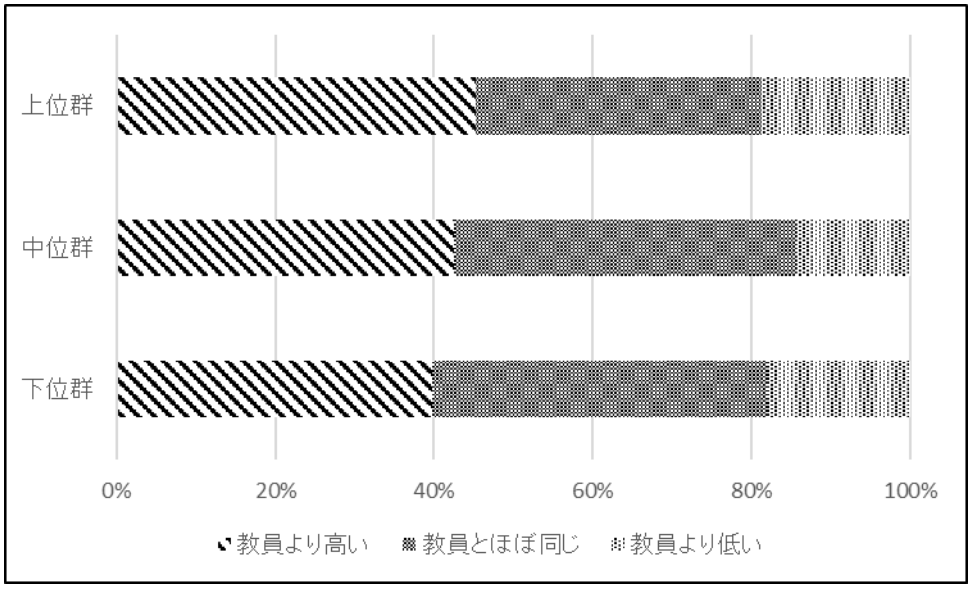


図 5. 学生による相互評価と教員による評価の評価の合計点の差

学生の評価傾向は，どの成績群においてもやや甘目になることがわかったが，上位群，中位群，下位群の間に差があるかをカイ 2 乗検定を用いて調べた。その結果を表 17 に示す。

表 17
カイ 2 乗検定の結果

	値	自由度	漸近有意確率（両側）	効果量（*Cramer's V*）	効果の大きさ
カイ 2 乗	1.831	4	.767	.06	無
尤度比	1.870	4	.760		
線型と線型による連関	.150	1	.698		
有効なケースの数	300				

図 5 と表 17 から，英語力の各成績群における評価者群（上位群，中位群，下位群）の間に有意差がなく，効果量も無い（*Cramer's V* =.06）ことがわかる。図 5 と表 17 の結果から，3 群の間に評価傾向の違いは見られないことから，学生の評価傾向は，英語力にかかわらず，教員の評価と比べるとやや甘目になると考えられる。

しかし，有意差がなくても，多重比較をすると有意差が出る場合があり，また，２群の間に差があるかを調べるために，カイ 2 乗検定を用いて多重比較を行った。さらに効果量を算出した。

表 18 に上位群と中位群におけるのカイ 2 乗検定の結果を示す。

表 18

カイ 2 乗検定の結果（上位群 VS 中位群）

	値	自由度	漸近有意確率（両側）	効果量（*Cramer's V*）	効果の大きさ
カイ 2 乗	1.423	2	.491	.08	無
尤度比	1.427	2	.490		
線型と線型による連関	.085	1	.771		
有効なケースの数	200				

表 18 から，上位群と中位群の間に有意差はなく，効果量も無い（*Cramer's V* =.08）ことから，上位群と中位群の間に差がないことがわかる。つまり，上位群と中位群の評価傾向の違いは見られない。

表 19 に中位群と下位群におけるカイ２乗検定の結果を示す。

表 19

カイ 2 乗検定の結果（中位群 VS 下位群）

	値	自由度	漸近有意確率（両側）	効果量（*Cramer's V*）	効果の大きさ
カイ 2 乗	.620	2	.733	.06	無
尤度比	.622	2	.733		
線型と線型による連関	.478	1	.489		
有効なケースの数	200				

表 19 から，中位群と下位群の間に有意差はなく，効果量も無い（*Cramer's V* =.06）ことから，中位群と下位群の間に差がないことがわかる。つまり，中位群と下位群の間に評価傾向の違いは見られない。

表 20 に上位群と下位群におけるにおけるカイ 2 乗検定の結果を示す。

表 20

カイ 2 乗検定の結果（上位群 VS 下位群）

	値	自由度	漸近有意確率（両側）	効果量（*Cramer's V*）	効果の大きさ
カイ 2 乗	.783	2	.676	.06	無
尤度比	.783	2	.676		
線型と線型による連関	.144	1	.704		
有効なケースの数	200				

表 20 から，上位群と下位群の間に有意差はなく，効果量も無い（*Cramer's V* =.06）ことから，上位群と下位群の間に差がないことがわかる。つまり，上位群と下位群の間に評価傾向の違いは見られない。

8. 全体の考察

6.1 では，評価者としての学生の評価力を調べるために，学生による相互評価と教員による評価との相関係数を求め分析した結果，すべての成績群の評価の合計点において，中程度の相関が得られたことから，どの成績群の学生も評価力が高いとは言えない。つまり，学生の英語力の高低と評価力の高低との間に関係はあまりないということがわかった。このことは，shimura (2006) の結果とも合致する。研究 1 で，教員による評価と学生による相互評価の相関係数が中程度になった要因の一つに，英語力が影響している可能性が考えられたが，, 英語力が高いことが必ずしも評価力が高いということを証明できなかった。

学生による相互評価と教員による評価の相関係数を評価項目別に見てみると，どの成績群においても，「準備」や「発表」に比べ，「リサーチ」と「オリジナリティ」の項目の相関が低くなる傾向があることがわかった。このことは，笠巻（2016）,笠巻（2018）の結果とも合致する。これは，「準備」や「発表」という評価項目は，その良し悪しが見た目にわかりやすく，評価経験の浅い学生にとっても評価しやすい項目であること，さらに，これらの項目には英語力が影響しにくいことが考えられる。一方，リサーチやオリジナリティという評価項目は内容に関わるものであることから，その評価には，当然その内容をよく理解することが求められる。つまりは，内容を十分に理解できるだけの英語力が必要になる。具体的には，発表内容を十分に理解できる英語力があれば，デリバリーと内容の両方を含めて，発表全体の良し悪しを判断できるものと考えられるが，内容を十分に理解できるだけの英語力がない評価者が，例えばデリバリーのみが上手い学生の発表を聴いた場合，内容の良し悪しに対する評価を考慮せず，デリバリーの良し悪しだけで，評価をしてしまう可能性が考えられる。「リサ

ーチ」の項目に対する下位群の学生の相関係数が，他の 2 群と比べ特に低いこともこのためと思われる。

しかし，発表内容の良し悪しを判断するには，英語力の他に，評価をする学生が発表内容の分野についてよく知っているか，つまりは，評価者の予備知識が大きく影響していると考えられる。つまり，英語力は高いがスキーマがなければ，発表内容を正しく理解できないことから評価力が低くなり，反対に英語力が高くなくても，発表内容に対するスキーマを十分に持っていれば，内容をよく理解できることから，発表の良し悪しをしっかりと判断できる可能性が考えられる。研究 2 では，発表内容に詳しい学生とそうでない学生が混在していたことも，他の評価項目に比べ，特に内容に関する項目の相関が低くなったものと考えられる。

6.2 では，学生の評価力が英語力の成績群間で差があるかを調べたが，学生の評価力は，英語力の上・中・下と同じ順にはならないことが確認された。そして，成績群の中で，中位群の評価力が一番高いことが判明した。本来，英語力が高いということは，発表内容を十分に理解できると予想されることから，上位群の評価力が一番高くなると予想されたが，そのような結果にはならなかった。すなわち，学生の英語力は学生の評価力とは関係がない可能性があると言えるだろう。

分析 2 では，分析 1 で，学生の評価力が高くないこと，そして，学生の英語力が学生の評価力とは関係がない可能性があるという結果から，学生の評価傾向を調べた。その結果，学生の評価傾向として，教員の評価と比べると甘目の評価を行う傾向があることがわかった。さらに成績群別に評価傾向を調べたが，成績群間に差がないことから，学生全体として，高めの点数をつける傾向があることがわかった。このことは，Chen & Warren (2005)，Fukazawa (2010)，笠巻 (2016)，笠巻 (2018) の結果とも合致する。つまり，学生による評価は，教員より甘めになることから，学生による相互評価の信頼性は低いと言えよう。

学生による評価が甘くなる傾向がある要因として，友達を評価することに対する気まずさから，いわゆる“お友達効果”と呼ばれる，心理的な要因が評価に影響を及ぼしている可能性があると思われる（Hanarahan & Issacs,2001）。また，田中（2017）は，他者との関係性を重要視する学生は評価が甘くなるなど，学生評価者の性格が相互評価にバイアスがかかってしまったり，正確性を欠いてしまう原因となりうるとも指摘している。このため，学生の評価の信頼性は低い（Freeman, 1995；笠巻, 2016；笠巻, 2018）と言えよう。

学生の英語力の他に，学生による評価に影響を及ぼした可能性があると考えられた原因として，以下のような点が考えられる。まず一つ目として，学生の評価経験の浅さが考えられる。大学 1 年生の前期の中間発表の時点では，学生の評価の経験はほとんどないと考えられることから，学生の評価力は非常に低いと思われる。評価経験の

有無は評価の信頼性や一貫性に影響すること（Weigle, 1994；山西, 2004；笠巻, 2018）や，評価を行う回数が増えることで，より正確な評価ができるようになる（Nakamura, 2002）ことから，実際に相互評価を行う前に，評価者トレーニングを実施することで，学生の評価経験の浅さを補えると考える。

二つ目に，研究 2 で使用している評価シートにある評価項目に対する判定基準の曖昧さが考えられる。この評価シートには，大きく 4 つの評価項目（準備，リサーチ，オリジナリティ，発表）があるが，発表には，発音，声の大きさ，速度，アイコンタクト，明瞭さという要素が含まれると記されている。例えば，学生が発表に対する評価に 3 を付けた場合，発表の項目の中のどの要素を見てその評価をしたのか，もしくはどの要素とどの要素の組み合わせで評価したのか，あるいはすべての要素から判断して評価したのかがわからない。学生は相互評価の際に，発表を聴きながら同時に評価を行っているため，果たして学生がすべての要素を組み入れて評価をしているとは考えにくい。また，学生によって組み合わせた要素も違っている可能性がある。これを改善にするためには，評価項目を，教員との相関係数が高い，中程度以上の項目だけに絞り，そしてそこに明確な判定基準を書いたルーブリックを使用することで，より明確に評価を行えるようになるのではないだろうかと考える。

最後に，評価には，上述したように発表内容に対する予備知識の有無や程度といった要因がからんで影響を及ぼしている可能性が考えられる。「プロジェクト発信型プログラム」では，学生が自分の興味・関心のあることに基づいてリサーチを行った成果を英語で発表することから，クラスメイトの発表内容は一人ひとり違うものになる。クラスメイト全員の発表内容について，学生が予備知識を持って評価を行うのは難しいと思われることから，評価項目に内容を加えるかどうか，また加えるならば，その判定基準は学生の予備知識に影響されないようにするなど見直す必要があると考える。

上述したように，評価には，ただ一つの要因だけではなく，他に複数の要因が評価に影響を及ぼすと考えられることから，学生の英語力の高低だけが学生の評価力に影響を及ぼすわけではないと言えよう。

注

1. TOEICIP テストの点数を学生が覚えている範囲のスコアを報告させている。

第5章　研究3:学生の「予備知識」が評価者としての学生の評価力に影響を及ぼすか？

1. はじめに

学生の評価力に影響を及ぼすとされる４つの要因（プレゼンテーション力, 英語力, 予備知識，評価者トレーニング）の一つである，学生の英語力が評価者としての学生の評価力に影響を及ぼすかを検証した第 4 章（研究 2）に続いて、研究 3 では，発表内容に対する予備知識の有無，またはその程度によって評価は変わるのか。また，それにより，教員による評価との相関は変わるのかについて検証する。

2. 研究 3 の目的

研究 3 では，評価者としての学生の評価力に影響を及ぼすと考えられる４つの要因のうち，学生の「予備知識」が評価力に影響を及ぼすかを調べるために，予備知識の程度別にみた学生の評価傾向を調べることを目的とする。

3. 研究 3 における指導と評価方法

3.1 指導

指導内容と授業の流れについては，p. 10, p 11 で述べたので，ここでは省略する。

3.2 評価方法

評価方法については，p. 11, p. 12 で述べたので，ここでは省略する。

4. 研究方法

4.1 参加者

研究 3 の参加者は, 滋賀県内の私立大学における筆者担当クラスの 1 年生計 61 名である。参加者の英語力は，TOEIC IP テストで平均 464 点であった。

4.2 手続き

クラスメイトの発表内容についてどのくらい知っていたかについて，学生の予備知識の程度を 2 から 10 の 5 段階に分け，予備知識 2 は全く知らない，予備知識 4 はほとんど知らない，予備知識 6 は詳しくは知らないがある程度は知っている，予備知識 8 はよく知っている，そして予備知識 10 はとても知っているとし，評価シートに該当する数字を記入させた。予備知識の程度の判断に迷った場合のために中間点を設けたため，原則的には 5 段階ではあるが，実質的には 10 段階としている。

学生の予備知識の高低の違いによって, 学生の評価傾向は変わるかを調べるために, 学生一人一人の発表に対する，教員による評価の点数と学生による相互評価の点数の

差を予備知識の程度別に調べた。

5. 分析

5.1 予備知識別に見た学生による評価傾向の分析

5.1.1 分析方法

学生の評価傾向を調べるために，中間発表に対する教員による評価の合計点の点数と学生による相互評価の合計点の点数の差を調べることにより分析した。予備知識の程度の違いによって差があるかをカイ 2 乗検定を用いて調べた。

5.1.2 分析結果と考察

学生の予備知識の各程度別に教員による評価と学生による評価の高低関係とその割合を表 1 と図 1 に示す。

表 1

予備知識レベルと教員による評価と学生による評価の高低関係とその割合

教員より低い	教員とほぼ同じ	教員より高い	(人数)合計	予備知識	教員より低い	教員とほぼ同じ	教員より高い	(割合)合計
7	8	3	18	10	38.89	44.44	16.67	100
17	37	39	93	8	18.28	39.78	41.94	100
66	78	120	264	6	25.00	29.55	45.45	100
75	85	92	252	4	29.76	33.73	36.51	100
34	41	50	125	2	27.20	32.80	40.00	100

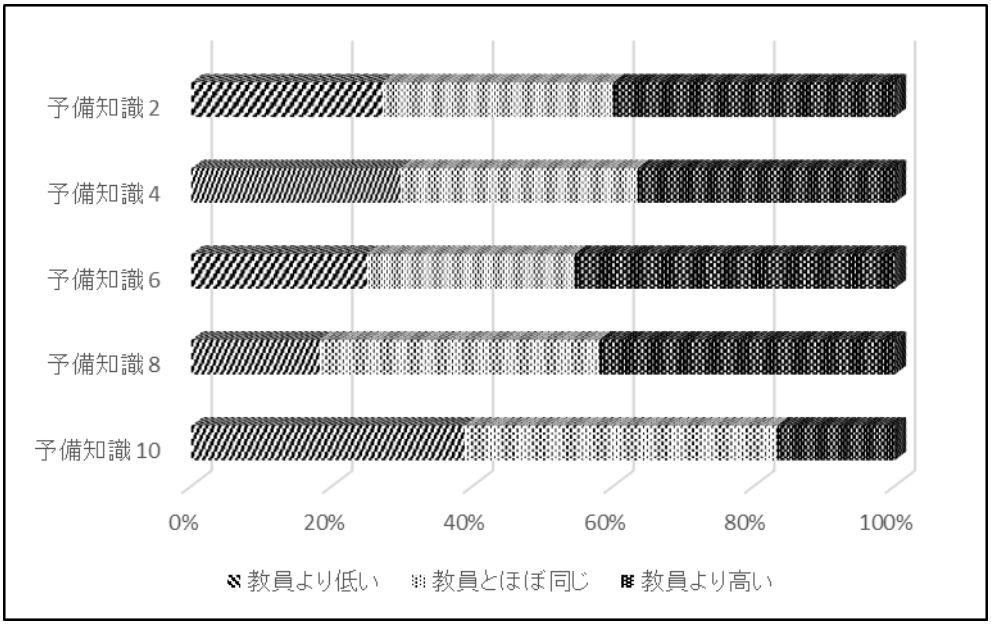


図 1. 予備知識レベル別に見た学生の評価傾向

表 1 と図 1 から，学生による評価は，予備知識 10 のグループを除いて，教員よりも高くつけていることがわかる。つまり，学生の評価傾向は，予備知識の程度に関わらず，やや甘目になる可能性が考えられる。

学生の評価傾向をさらに詳しく調べるため, 予備知識 2, 4, 6, 8, 10 の間に差があるかをカイ 2 乗検定を用いて調べた。その結果を表 2 に示す。

表 2

カイ 2 乗検定の検定結果・効果量

	値	自由度	漸近有意確率 (両側)	効果量 (*Cramer's V*)	効果の大きさ
Pearson のｶｲ 2 乗	25.713	8	.001	.131	小
尤度比	27.920	8	.000		
線型と線型による連関	4.419	1	.036		
有効なｹｰｽの数	501				

図 1 と表 2 から，予備知識の各レベル群（2, 4, 6, 8, 10）の間に差があり，効果量が小であることがわかる。この結果から，予備知識の程度によって，学生の評価傾向には違いがあることがわかった。

5.2　予備知識の高いグループ VS 予備知識の低いグループの評価傾向の分析

5.2.1　分析方法

5.1 で，学生の予備知識の程度によって，評価傾向に差があることがわかったが，どこに差があるかをさらに詳しく調べるため, 予備知識 8 のグループと予備知識 10 のグループを併せて「予備知識の高いグループ」とし，予備知識 2 と 4 のグループを併せて「予備知識の低いグループ」として，両グループの間に評価傾向の差があるかを調べた。検定方法は，5.1.1 と同じである。

5.2.2　分析結果と考察

学生の予備知識の高低別に教員による評価と学生による評価の高低関係とその割合を表 3 と図 2 に示す。

表 3

予備知識の高低と教員による評価と学生による評価の高低関係とその割合

教員より低い	教員とほぼ同じ	教員より高い	(人数) 合計	予備知識	教員より低い	教員とほぼ同じ	教員より高い	(割合) 合計
24	45	42	111	高	21.6	40.5	37.8	100
109	126	142	377	低	28.9	33.4	37.7	100

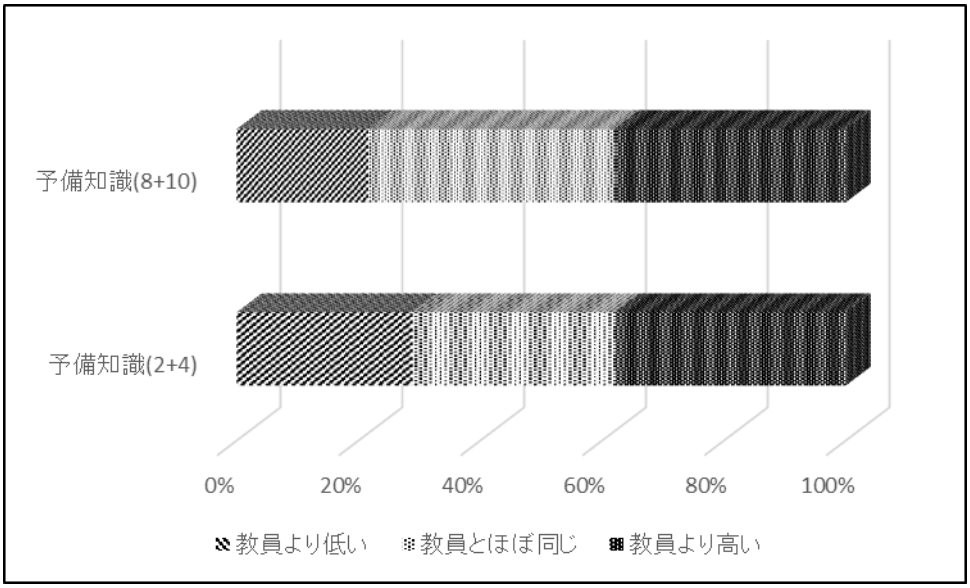


図 2. 予備知識の高低別に見た学生の評価傾向

予備知識の高いグループと低いグループとの間に差があるかをカイ 2 乗検定を用いて調べた。その結果を表 4 に示す。

表 4
カイ 2 乗検定の検定結果・効果量

	値	自由度	漸近有意確率 (両側)	効果量 (*Cramer's V*)	効果の大きさ
Pearson のｶｲ 2 乗	1.821	2	.402	.062	無
尤度比	1.825	2	.401		
線型と線型による連関	.380	1	.538		
有効なｹｰｽの数	201				

表 3, 表 4 および図 2 から, 予備知識の高いグループと低いグループとの間に有意差はなく，効果量もないことから，予備知識の高いグループと低いグループとの間に差がないことがわかる。つまり，予備知識の高いグループと低いグループとの間に評価傾向の違いは見られない。

予備知識の高いグループと低いグループとの間に有意差はなかったが，多重比較をすると有意差が出る場合があり、また、各予備知識グループの評価傾向の差がどの程度の大きさであるかを調べるために、カイ 2 乗検定を用いて，多重比較を行った。さらに効果量を算出した。

5.3 予備知識グループ間の評価傾向の分析

5.3.1 分析方法

各予備知識グループの間に評価傾向の差があるかを調べた。検定方法は, 5.1.1, 5.2.1 と同じとする。

5.3.2 分析結果と考察

予備知識 2 グループと予備知識 4 グループとの間に差があるかをカイ 2 乗検定を用いて調べた。その結果を表 5 と 6 に示す。

表 5

予備知識 2 vs 予備知識 4 の教員による評価と学生による評価の高低関係とその割合

教員より低い	教員とほぼ同じ	教員より高い	(人数)) 合計	予備知識	教員より低い	教員とほぼ同じ	教員より高い	(割合)) 合計
75	85	92	252	4	29.76	33.73	36.51	100
34	41	50	125	2	27.20	32.80	40.00	100

表 6

カイ 2 乗検定結果（予備知識 2 VS 予備知識 4）・効果量

	値	自由度	漸近有意確率（両側）	効果量（*Cramer's V*）	効果の大きさ
カイ 2 乗	.285	2	.867	.027	無
尤度比	.285	2	.867		
線型と線型による連関	.280	1	.596		
有効なケースの数	201				

表 5 と 6 から, 予備知識 2 のグループと予備知識 4 のグループの間に有意差はなく, 効果量もないことがわかる。つまり，予備知識 2 のグループと予備知識 4 のグループの間に差がないことから，発表内容について全く知らない（予備知識 2）程度とほとんど知らない（予備知識 4）程度の間には評価傾向の違いはないと言えよう。

次に，予備知識 2 グループと予備知識 6 グループとの間に差があるかをカイ 2 乗検定を用いて調べた。その結果を表 7 と 8 に示す。

表 7

予備知識 2 vs 予備知識 6 の教員による評価と学生による評価の高低関係とその割合

教員より低い	教員とほぼ同じ	教員より高い	(人数) 合計	予備知識	教員より低い	教員とほぼ同じ	教員より高い	(割合) 合計
66	78	120	264	6	25.00	29.55	45.45	100
34	41	50	125	2	27.20	32.80	40.00	100

表 8

カイ 2 乗検定結果（予備知識 2 VS 予備知識 6）・効果量

	値	自由度	漸近有意確率（両側）	効果量（*Cramer's V*）	効果の大きさ
カイ 2 乗	.514	2	.773	.036	無
尤度比	.514	2	.773		
線型と線型による連関	.371	1	.543		
有効なケースの数	200				

表 7 と 8 から, 予備知識 2 のグループと予備知識 6 のグループの間に有意差はなく, 効果量もないことがわかる。つまり，予備知識 2 のグループと予備知識 6 のグループの間に差がないことから，発表内容について全く知らない（予備知識 2）程度と詳しくは知らないが，ある程度は知っている（予備知識 6）程度の間には評価傾向の違いはないと言えよう。

次に，予備知識 2 グループと予備知識 8 グループとの間に差があるかをカイ 2 乗検定を用いて調べた。その結果を表 9 と 10 に示す。

表 9

予備知識 2 vs 予備知識 8 の教員による評価と学生による評価の高低関係とその割合

教員より低い	教員とほぼ同じ	教員より高い	(人数) 合計	予備知識	教員より低い	教員とほぼ同じ	教員より高い	(割合) 合計
17	37	39	93	8	18.28	39.78	41.94	100
34	41	50	125	2	27.20	32.80	40.00	100

表 10

カイ 2 乗検定結果（予備知識 2 VS 予備知識 8）・効果量

	値	自由度	漸近有意確率（両側）	効果量（*Cramer's V*）	効果の大きさ
カイ 2 乗	2.520	2	.284	.108	小
尤度比	2.533	2	.282		
線型と線型による連関	1.002	1	.317		
有効なケースの数	200				

表 9 と 10 から, 予備知識 2 のグループと予備知識 8 のグループの間に有意差はない

が，効果量が小（r = .10）で差が見られた。つまり，予備知識 2 のグループと予備知識 8 のグループの評価傾向に違いが少しある可能性が考えられる。この結果から，発表内容について全く知らない（予備知識 2）程度とよく知っている（予備知識 8）程度では，予備知識の程度の差が大きくなることから，評価者の予備知識の多少によって評価傾向に違いが出る可能性あると考えられる。

次に, 予備知識 2 グループと予備知識 10 グループとの間に差があるかをカイ 2 乗検定を用いて調べた。その結果を表 11 と 12 に示す。

表 11

予備知識 2 vs 予備知識 10 の教員による評価と学生による評価の高低関係とその割合

教員より低い	教員とほぼ同じ	教員より高い	(人数) 合計	予備知識	教員より低い	教員とほぼ同じ	教員より高い	(割合) 合計
7	8	3	18	10	38.89	44.44	16.67	100
34	41	50	125	2	27.20	32.80	40.00	100

表 12

カイ 2 乗検定結果（予備知識 2 VS 予備知識 10）・効果量

	値	自由度	漸近有意確率（両側）	効果量（*Cramer's V*）	効果の大きさ
カイ 2 乗	13.034	2	.001	.302	中
尤度比	13.321	2	.001		
線型と線型による連関	9.942	1	.002		
有効なケースの数	200				

表 11 と 12 から，予備知識 2 のグループと予備知識 10 のグループの間に差があり，効果量が中（r = .30）であることがわかる。つまり，予備知識 2 のグループと予備知識 10 のグループの評価傾向に違いがある可能性が考えられる。この結果から，発表内容について全く知らない（予備知識 2）程度ととても知っている（予備知識 10）程度では，予備知識の程度の差がさらに大きくなることから，評価者の予備知識の多少によって評価傾向に違いが出る可能性があると考えられる。

次に，予備知識 4 グループと予備知識 6 グループとの間に差があるかをカイ 2 乗検定を用いて調べた。その結果を表 13 と 14 に示す。

表 13

予備知識 4 vs 予備知識 6 の教員による評価と学生による評価の高低関係とその割合

教員より低い	教員とほぼ同じ	教員より高い	(人数) 合計	予備知識	教員より低い	教員とほぼ同じ	教員より高い	(割合) 合計
66	78	120	264	6	25.00	29.55	45.45	100
75	85	92	252	4	29.76	33.73	36.51	100

表 14

カイ 2 乗検定結果（予備知識 4 VS 予備知識 6）・効果量

	値	自由度	漸近有意確率（両側）	効果量（*Cramer's V*）	効果の大きさ
カイ 2 乗	1.480	2	.477	.054	無
尤度比	1.482	2	.477		
線型と線型による連関	1.287	1	.257		
有効なケースの数	201				

表 13 と 14 から，予備知識 4 のグループと予備知識 6 のグループの間に有意差はなく，効果量もないことがわかる。つまり，予備知識 4 のグループと予備知識 6 のグループの間に差がないことから，これらのグループ間における評価傾向に違いはないと言えよう。これは，発表内容に対してほとんど知らない（予備知識 4）と詳しくは知らないが，ある程度は知っている（予備知識 6）とでは，予備知識の程度に差がほとんどないことから，評価傾向に違いが見られなかったと考えられる。

次に，予備知識 4 グループと予備知識 8 グループとの間に差があるかをカイ 2 乗検定を用いて調べた。その結果を表 15 と 16 に示す。

表 15

予備知識 4 vs 予備知識 8 の教員による評価と学生による評価の高低関係とその割合

教員より低い	教員とほぼ同じ	教員より高い	(人数) 合計	予備知識	教員より低い	教員とほぼ同じ	教員より高い	(割合) 合計
17	37	39	93	8	18.28	39.78	41.94	100
75	85	92	252	4	29.76	33.73	36.51	100

表 16

カイ 2 乗検定結果（予備知識 4VS 予備知識 8）・効果量

	値	自由度	漸近有意確率（両側）	効果量（*Cramer's V*）	効果の大きさ
カイ 2 乗	3.798	2	.150	.105	小
尤度比	3.831	2	.147		
線型と線型による連関	2.396	1	.122		
有効なケースの数	201				

表 15 と 16 から，予備知識 4 のグループと予備知識 8 のグループの間に有意差はないが，効果量が小（r = .10）で差が見られた。つまり，予備知識 4 のグループと予備知識 8 のグループの評価傾向に違いが少しある可能性が考えられる。これは，発表内容に対してほとんど知らない（予備知識 4）とよく知っている（予備知識 8）では，予備知識の程度の差が大きくなるため, 評価傾向に違いが見られた可能性が考えられる。

次に, 予備知識 4 グループと予備知識 10 グループとの間に差があるかをカイ 2 乗検定を用いて調べた。その結果を表 17 と 18 に示す。

表 17

予備知識 4 vs 予備知識 10 の教員による評価と学生による評価の高低関係とその割合

教員より低い	教員とほぼ同じ	教員より高い	(人数) 合計	予備知識	教員より低い	教員とほぼ同じ	教員より高い	(割合) 合計
7	8	3	18	10	38.89	44.44	16.67	100
75	85	92	252	4	29.76	33.73	36.51	100

表 18

カイ 2 乗検定結果（予備知識 4 VS 予備知識 10）・効果量

	値	自由度	漸近有意確率（両側）	効果量（*Cramer's V*）	効果の大きさ
カイ 2 乗	9.859	2	.007	.194	小
尤度比	10.045	2	.007		
線型と線型による連関	6.901	1	.009		
有効なケースの数	201				

表 17 と 18 から, 予備知識 4 のグループと予備知識 10 のグループの間に有意差はな

いが，効果量が小（r = .19）であることがわかる。つまり，予備知識 4 のグループと予備知識 10 のグループの評価傾向に違いが少しある可能性が考えられる。これは，発表内容に対してほとんど知らない（予備知識 4）ととても知っている（予備知識 10）の間における予備知識の程度の差がさらに大きくなることから，評価傾向に少し違いが見られたと考えられる。

次に，予備知識 6 グループと予備知識 8 グループとの間に差があるかをカイ 2 乗検定を用いて調べた。その結果を表 19 と 20 に示す。

表 19

予備知識 6 vs 予備知識 8 の教員による評価と学生による評価の高低関係とその割合

教員より低い	教員とほぼ同じ	教員より高い	(人数) 合計	予備知識	教員より低い	教員とほぼ同じ	教員より高い	(割合) 合計
17	37	39	93	8	18.28	39.78	41.94	100
66	78	120	264	6	25.00	29.55	45.45	100

表 20

カイ 2 乗検定結果（予備知識 6 VS 予備知識 8）・効果量

	値	自由度	漸近有意確率（両側）	効果量（*Cramer's V*）	効果の大きさ
カイ 2 乗	2.672	2	.263	.087	無
尤度比	2.682	2	.262		
線型と線型による連関	0.132	1	.716		
有効なケースの数	200				

表 19 と 20 から，予備知識 6 のグループと予備知識 8 のグループの間に有意差はなく，効果量もないことがわかる。つまり，予備知識 6 のグループと予備知識 8 のグループの間に差がないことから，発表内容に対して詳しくは知らないが，ある程度は知っている（予備知識 6）程度とよく知っている（予備知識 8）程度の評価傾向に違いはないと言えよう。これは，予備知識 6 のグループと予備知識 8 のグループとでは，予備知識の程度の差が小さいことから, 評価傾向に違いが見られなかったと考えられる。

次に, 予備知識 6 グループと予備知識 10 グループとの間に差があるかをカイ 2 乗検定を用いて調べた。その結果を表 21 と 22 に示す。

表 21

予備知識 6 vs 予備知識 10 の教員による評価と学生による評価の高低関係とその割合

教員より低い	教員とほぼ同じ	教員より高い	(人数)合計	予備知識	教員より低い	教員とほぼ同じ	教員より高い	(割合)合計
7	8	3	18	10	38.89	44.44	16.67	100
66	78	120	264	6	25.00	29.55	45.45	100

表 22

カイ 2 乗検定結果（予備知識 6 VS 予備知識 10）・効果量

	値	自由度	漸近有意確率（両側）	効果量（*Cramer's V*）	効果の大きさ
カイ 2 乗	18.356	2	.000	.255	小
尤度比	18.867	2	.000		
線型と線型による連関	13.932	1	.000		
有効なケースの数	200				

表 21 と 22 から，予備知識 6 のグループと予備知識 10 のグループの間に差があり，効果量が小であることがわかる。つまり，発表内容に対して詳しくは知らないが，ある程度は知っている（予備知識 6）程度と，とてもよく知っている（予備知識 10）程度の間には予備知識の程度の差があるため, 評価傾向に違いが見られたと考えられる。

次に, 予備知識 8 グループと予備知識 10 グループとの間に差があるかをカイ 2 乗検定を用いて調べた。その結果を表 23 と 24 に示す。

表 23

予備知識 8 vs 予備知識 10 の教員による評価と学生による評価の高低関係とその割合

教員より低い	教員とほぼ同じ	教員より高い	(人数)合計	予備知識	教員より低い	教員とほぼ同じ	教員より高い	(割合)合計
7	8	3	18	10	38.89	44.44	16.67	100
17	37	39	93	8	18.28	39.78	41.94	100

表 24

カイ２乗検定結果（予備知識 8 VS 予備知識 10）・効果量

	値	自由度	漸近有意確率（両側）	効果量（*Cramer's V*）	効果の大きさ
カイ 2 乗	18.521	2	.000	.408	中
尤度比	19.048	2	.000		
線型と線型による連関	18.153	1	.000		
有効なケースの数	200				

表 23 と 24 から，予備知識 8 のグループと予備知識 10 のグループの間に差があり，効果量が中であることがわかる。つまり，予備知識 8 のグループと予備知識 10 のグループの評価傾向に違いがある可能性が考えられる。しかし，発表内容に対してよく知っている（予備知識 8）程度ととても知っている（予備知識 10）程度の間における予備知識の程度の差が小さいにもかかわらず，両グループの間に差が見られたことは，これまでの結果と比べると，予備知識 8 のグループと予備知識 10 のグループの間における評価傾向に違いがあるというよりは，偶然差が出た可能性が考えられる。

6. 全体の考察

学生の評価傾向として，教員の評価と比べると高めの点数をつける傾向があることがわかった。このことは，Chen & Warren (2005)，Fukazawa (2010)，笠巻 (2016),笠巻 (2018)，そして，研究 1 と研究 2 の結果とも合致する。つまり，学生による評価は甘くなることが研究 3 でも確認された。学生による評価が甘くなる要因として，Hanarahan & Issacs (2001)と田中 (2017)は，クラスメイトを評価することに対して，心理的な要因が評価に影響を及ぼしている可能性があることを指摘している。

5.1.で，学生が持つ，クラスメイトの発表内容に対する予備知識の程度の違いによって，学生の評価傾向には違いがある可能性があることがわかったことから，5.2 では，予備知識の程度のどこに差があるかを調べ，予備知識の高いグループと低いグループとに分けて比較し分析した結果，予備知識の高低の間には評価傾向の違いは見られなかった。そこで，5.3 では，各予備知識グループの間における評価傾向に違いがあるかを調べたところ，予備知識の程度の差が大きいほど，評価傾向に差が見られることがわかった。これらの結果から，予備知識の多少が学生の評価傾向に少しの影響を及ぼす可能性が考えられた。

研究 3 では，予備知識が高いほど評価は厳しめになり，反対に予備知識が低ければ甘目の評価になることが予想されたが，予備知識の程度の違いによる明確な評価傾向を知ることができなかった。

このような結果になった原因として，次のことが考えられる。一つ目は，発表内容に対する評価者の予備知識というのは，主に発表内容の理解度と関係していることが考えられる。つまり，良く知っている内容と，あまりよく知らない内容とでは，発表内容の理解度は変わってくることから，評価も変わるはずである。しかし，発表内容の理解には，ただ単に評価者の予備知識だけではなく，英語による発表であることからも，内容を十分に理解できるだけでの英語力も関係していると考えられる。これが，評価者の予備知識だけに焦点を当てた研究 3 の分析では，明確な評価傾向が得られなかった原因と考えられる。

二つ目の理由として，学生の評価経験の浅さが考えられる。おそらく学生は，予備知識の有無や程度の違いで，評価方法を変えた可能性が考えられる。その変更の仕方は一貫しているか，していないかを測ることはできないが，学生の評価経験の浅さから考えると，学生の評価方法に一貫性があるとは考えにくい。評価経験が浅く，まだ評価力の低い学生が一貫した評価を行えるようにするには，評価者トレーニングを実施することが必要であろう。評価経験の有無は評価の信頼性や一貫性に影響すること（Weigle, 1994；山西，2004；笠巻，2018）や，評価を行う回数が増えることで，より正確な評価ができるようになる（Nakamura, 2002）ことから，実際に相互評価を行う前に，評価者トレーニングを実施することで，学生の評価経験の浅さを補えると考える。

さらに三つ目の理由として，相互評価の実施方法が考えられる。研究 3 では，学生に，5 段階に分けられた予備知識の程度を，他の評価項目と同じく評価シートに記入させたが，クラスメイトの発表の評価をしながら，同時に発表内容に対する自分の予備知識も判断するというのは，評価経験の浅い学生にとってどれほど難しかったことは想像に難くない。「プロジェクト発信型プログラム」では，学生の発表内容は一人一人違うことに加え，一度の発表で約 10 名程度の発表を評価することから，上記のような判断をして評価に臨むのは学生にとっては負担が大きかった可能性がある。つまり，そのような状況での学生の評価は信頼性が低いことが考えられる。この点においては，評価項目を発表の仕方ではなく発表の内容に絞るなどして見直すことや，また，クラスメイトの発表の評価が終わってから，予備知識について尋ねるなど，相互評価の実施方法も見直す必要がある。

第6章　研究4：「評価者トレーニング」が評価者としての学生の評価力に影響を及ぼすか？

1. はじめに

学生の評価力に影響を及ぼすとされる 4 つの要因（プレゼンテーション力, 英語力, 予備知識，評価者トレーニング）の一つである，学生の予備知識が評価者としての学生の評価力に影響を及ぼすかを検証した第 5 章（研究 3）では，予備知識の多少が学生の評価傾向に少しの影響を及ぼす可能性が考えられたが，予備知識の高低によって学生の評価傾向に違いが現れる明確な要因を突き止めることができなかった。

研究 4 は，上述の要因の一つである，評価者トレーニングの有無によって，プレゼンテーションに対する学生による相互評価と教員による評価との相関が変わるかを調べ，評価者トレーニングが学生の評価力に影響を及ぼすかを検証する。

2. 評価者トレーニングに関する先行研究

評価者トレーニングとは，スピーキングやライティングなどのテストにおいて，受験者のパフォーマンスの評価に採点者側の主観が入る場合，複数の採点者が一貫した基準で評価できるようになるために行われるトレーニングのことであり（静他, 2002），スピーキングテスト，またパフォーマンステストにおいて，重要な要素であると言われている（Taylor & Galaczi, 2011, 望月他, 2015）。

評価者トレーニングを行うことで，学生による相互評価と教員による評価の相関が高くなるとの報告（Fukazawa, 2010; Okuda & Otsu, 2010; Patri, 2002）や，評価の点数より，コメントによる評価力を高めたり，ミスフィットする評価者を減らすのに効果があるとの報告もある（Saito, 2008）。

また，評価者トレーニングは，複数の評価者における評価者間信頼性より評価者内信頼性といった，一貫性に効果があるとも報告されている（Taylor & Galaczi, 2011; McNamara, 1996）。

そして，トレーニングの方法として，Hughes (1989) は，ルーブリックと過去の学生の発表動画を用いて，明示的に指導していくことが有効であるとしている。

3. 研究 4 の目的

研究 4 では，学生の評価に影響を及ぼす 4 つの要因の一つである，評価者トレーニングの有無によって，プレゼンテーションに対する学生による相互評価と教員による評価との相関が変わるかを調べ，評価者トレーニングが学生の評価力に影響を及ぼすかを検証すること目的とし，分析 1 と 2 を行った。

分析 1. 評価者トレーニングは，学生の評価力に影響を及ぼすか。また，どの評価項目に影響を及ぼすか。

分析 2. 評価者トレーニングの回数は，学生の評価力に影響を及ぼすか。また，どの評価項目に影響を及ぼすか。

4. 研究 4 における指導と評価方法

4.1 指導

指導内容と授業の流れについては，p. 10, p 11 で述べたので，ここでは省略する。

4.2 評価方法

評価方法については，p. 11, p. 12 で述べたので，ここでは省略する。

評価項目である，「準備」，「リサーチ」，「オリジナリティ」と「発表」は，「プロジェクト発信型プログラム」では統一されており，評価者トレーニングを行うために，これらの項目を細分化して，「内容」，「オリジナリティ」，「アイコンタクト」，「スライド」，「発音」，「ポーズの位置」とした。

また，「プロジェクト発信型プログラム」では，事前に評価トレーニングなどは行っていないが，研究 4 では評価トレーニングを行った後に，学生の相互評価を行った。，相互評価の際に用いる評価シートは，上述の評価項目をもとに，筆者が作成したルーブリックに基づいて作成し直したものを使用した（資料 2)。

5. 研究方法

5.1 参加者

研究 4 の参加者は，滋賀県内の私立大学における筆者担当クラスの 1 年生 4 クラス計 61 名である。参加者の英語力は，TOEIC IP テスト平均 464 点である。

5.2 分析対象者

参加者 61 名のうち，本格的な評価者トレーニングを行う 2 クラス（29 名）を詳細グループとし，簡易的な評価者トレーニングを行う 2 クラス（32 名）を簡易グループと設定した。1 回目の評価者トレーニング後の中間発表における学生による評価の合計の点数と教員による評価の合計の点数の相関係数が同じになるようにして，両グループの評価力を等質にし，詳細グループ（12 名），簡易グループ（12 名）の計 24 名を分析対象者とした。

表 1 と表 2 に「合計」における教員による評価と学生による評価の相関係数の記述統計量と検定結果を示す。

表 1

「合計」の記述統計量

	n	*M*	*SD*	最大値	最小値
詳細グループ	12	.383	.277	.799	-.101
簡易グループ	12	.383	.271	.788	-.074

表 2

「合計」のマン・ホイットニーの *U* 検定結果・効果量

	U	*W*	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
詳細グループ							
VS	70.5	148.5	-.087	.932	-.018	なし	
簡易グループ							

表 1 と 2 から，詳細グループと簡易グループの間にはは，評価力において，差がないことがわかる。つまり，両グループの評価力は等質であると言える。

5.3 研究 4 における評価者トレーニング

5.3.1 評価者トレーニングを行う前の準備

5.3.1.1 ルーブリック

研究 4 では，上述の評価項目をもとに，筆者がルーブリックを作成した。「プロジェクト発信型英語プログラム」では，評価項目の「オリジナリティ」を熱意と説得力と解釈しているが，これらは数値で判断することが難しいため，学生にとって判定基準をわかりやすくするために，発表内容を初めて聞いた内容かどうかに変更した。

図 1 に評価者トレーニングで使用したルーブリックを示す。

注意事項	P2最終発表　Rubrics						
・10段階評価							
・迷ったら、間の数字		内容		発表			
・小数は使わない	スコア	内容	オリジナリティ	アイコンタクト	スライド	発音	ポーズ
ほぼ完璧なレベル→	10	とてもよくわかった。	ほとんど初めて聞いた内容だった。	完全にオーディエンスに向かって発表している	ポイントが明確で、図やグラフなどが非常に効果的に使われていた。	とても英語らしく聞こえる	適切な位置にポーズが入っていた
	8	ほとんどわかった。	新しく知った内容が多かった。	ほぼオーディエンスに向かって話している	たまにわかりにくい箇所があるが、ほぼ伝えたい内容がわかる	たまにカタカナ英語が聞かれるが、ほとんど英語らしく聞こえる	ほぼ適切な位置にポーズが入っていた
	6	わかったところと、わからなかったところが半々。	知っている内容と知らなかった内容が半々	オーディエンスを見て話している時と、見ていない時と半々	わかりやすいスライドとわかりにくスライドが半々	カタカナ英語と英語らしく聞こえるのと半々	適切な位置にポーズが入っていたり、いなかったりした
	4	あまりよくわからなかった。	ほとんどすでに知っている内容だった	オーディエンスに向かって話していない方が多い。	図やグラフなども使われておらず、伝えたいポイントが不明確である。	カタカナ英語で発音していることの方が多く、あまり英語らしく聞こえない	適切な位置にポーズがあまり入っていない
実質上の最低レベル→	2	全然わからなかった。	全て知っている内容だった	ほとんどオーディエンスに向かって話していない。	文章が多すぎて、わかりにくい	完全なカタカナ英語で、英語らしく聞こえない	ポーズを入れる箇所が間違っていて、内容を理解できない。
	0	発表者はプレゼンテーションをしていない					

図１．評価者トレーニングで使用したルーブリック

5.3.1.2　ルーブリックの判定基準を説明するための音声およびスライドのサンプル

筆者が，各判定基準のレベルに応じて，「発音」と「速度」のデモ音声を作成した。例となる「スライド」も同様に作成した。図 2 から図 4 にそのサンプルを示す。

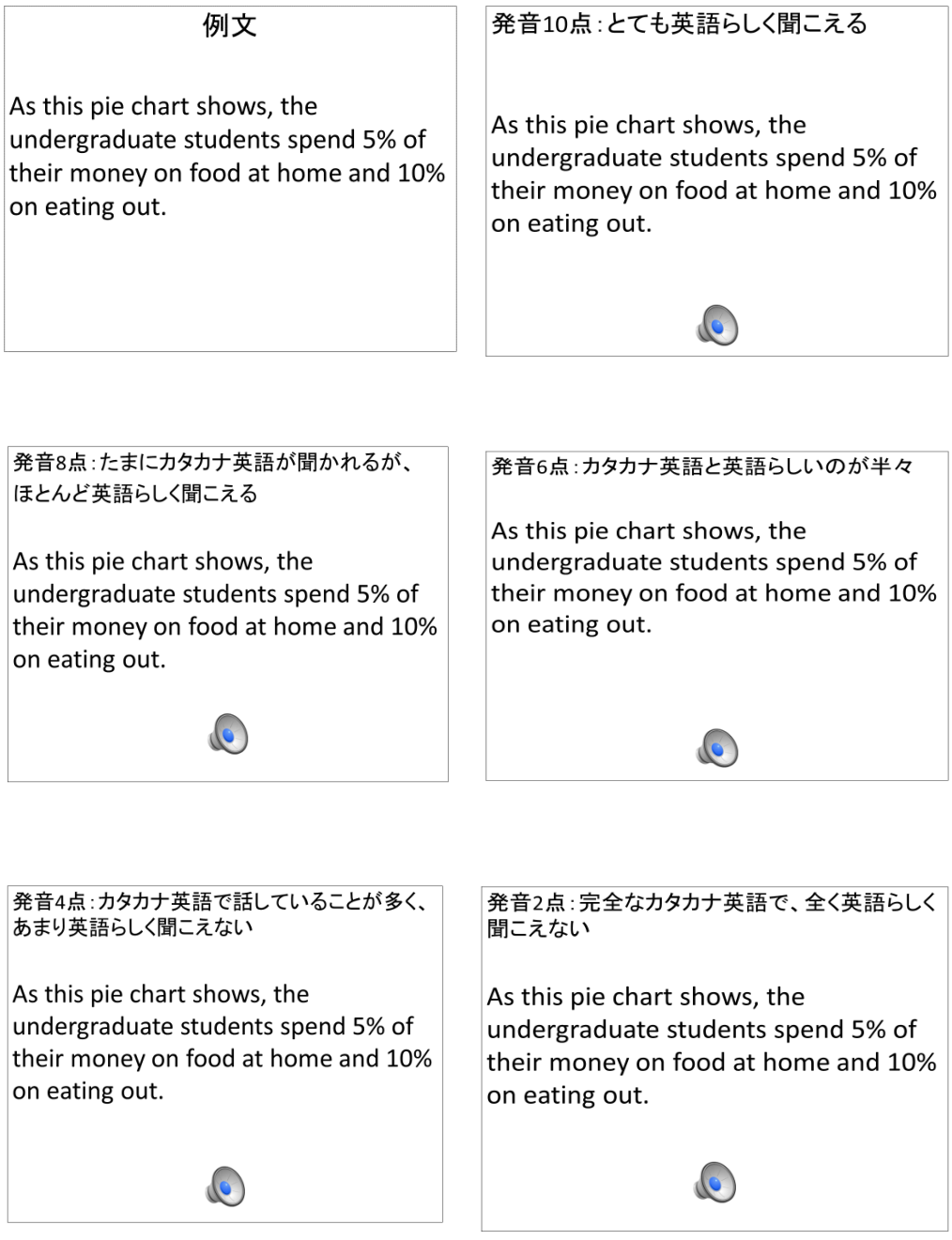


図 2. 評価者トレーニングで使用した音声サンプル（発音）

例文

As this pie chart shows, the undergraduate students spend 5% of their money on food at home and 10% on eating out.

ポーズ10点
適切な位置にポーズが入っていた

As this pie chart shows, /
the undergraduate students/
spend 5% of their money/
on food at home /
and 10% /on eating out.

ポーズ8点
ほぼ適切な位置にポーズが入っていた

As this pie chart/shows, /
the undergraduate students/
Spend/ 5% of /their money/
on food /at home /
and 10% on/ eating out.

ポーズ6点
適切な位置にポーズが入っていたり、いなかったりした

As this pie chart shows, /
the undergraduate students
spend / 5% of their money/
on food at home/
and /10% on eating out.

ポーズ4点
適切な位置にポーズがあまり入っていない

As this pie chart shows, /
the undergraduate students
spend 5% of their money/
on food /at home
and 10% on eating out.

ポーズ2点
ポーズを入れる箇所が間違っていて、内容が理解できない

As this pie chart/ shows,
the /undergraduate students
spend 5% of their/ money
on /food at/ home
and 10% on/ eating/ out.

図 3. 評価者トレーニングで使用した音声サンプル（ポーズ）

例文

Ritsumeikan University

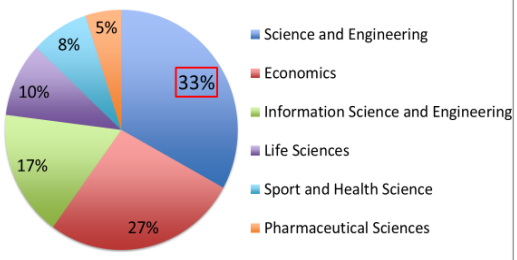
- It has fifteen thousands students.
- It has six department.
- The percentage of students in each faculty

Science and Engineering (33%)
Economics (27%)
Information Science and Engineering (17%)
Life Sciences (10%)
Sport and Health Science (8%)
Pharmaceutical Sciences (5%)

スライド10点：とても見やすく，伝えたい内容がよくわかる

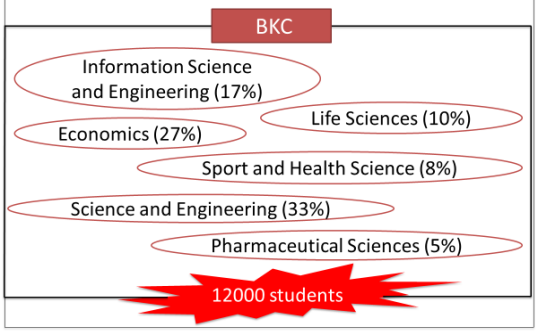
Ritsumeikan University

The percentage of students in each faculty (BKC)



スライド8点：たまにわかりにくい箇所があるが，ほぼ伝えたい内容がわかる

Ritusmeikan University



スライド6点：わかりやすいスライドとわかりにくいスライドが半々

Ritsumeikan University

- There are 12,000 students in BKC.
- The percentage of students in each faculty are listed below.

Science and Engineering (33%)
Economics (27%)
Information Science and Engineering (17%)
Life Sciences (10%)
Sport and Health Science (8%)
Pharmaceutical Sciences (5%)

スライド4点：何を伝えたいのか，あまりよくわからない

Ritsumeikan University

- There are 12000 students in BKC.
- 33% of students are in faculty of Science and Engineering.
- 27% of students are in faculty of Economics.
- 17% of students are in faculty of Information Science and Engineering.
- 10% of students are in faculty of Life Sciences.
- 8% of students are in faculty of Sport and Health Science.
- 7% of students are in faculty of Pharmaceutical Sciences.

スライド2点：何を伝えたいのか全くわからない

There are faculty of Science and Engineering, faculty of Economics, faculty of Information Science and Engineering, faculty of Life Sciences, faculty of Sport and Health Science, faculty of Pharmaceutical Sciences.

図 4. 評価者トレーニングで使用したスライドのサンプル

5.3.1.3　評価者トレーニングで使用するサンプルビデオ

筆者が過去に指導した学生に，研究 4 の主旨を説明し，自身の発表ビデオをサンプルとして使用させてもらう承諾を得た。賛同してくれた学生の発表ビデオの中から，判定基準に照らし合わせて，高いレベルの発表，中程度の発表，そして低いレベルの発表を行った学生を選び，サンプルビデオとしてトレーニングに使用した。

5.3.2　実施手順

研究 4 では，発表は中間発表と最終発表の２回であり，各発表の前に評価者トレーニングを行った。詳細グループに対し，中間発表の１週間前の授業時に１回目，最終発表の１週間前の授業時に２回目の評価者トレーニングを，下記の手順で行った。

1. 学生にルーブリックと評価シートを配布。
2. 上記のサンプルを用いて，ルーブリックに書かれている評価項目および評価基準の説明を行う。「内容」と「オリジナリティ」については，口頭による説明を行い，「アイコンタクト」については，筆者が実演した。
3. サンプル動画の中から一人（A 君）のビデオを見せて，評価させた後，評価シートを回収する。
4. 「内容」，「スライド」，「アイコンタクト」，「発音」，「ポーズ」の項目の判定基準の各レベルのデモを聞かせ，説明を行う。
5. サンプル動画の中から二人（B 君，C 君）のビデオを見せて，評価させる。
6. 教員から B 君，C 君の問題点あるいは改善点についてコメントをする。
7. 評価シートを回収する。
8. 新たに評価シートを配布し，再度，二人（B 君，C 君）のビデオを見せて，評価させる。
9. 評価シートを回収し，答え合わせを行う。
10. 新たに評価シートを配布し，再度，A 君のビデオを見せて，評価させる。
11. 評価シートを回収し，答え合わせを行う。

詳細トレーニングに要した時間は約 45 分であった。

簡易グループにおいては，中間発表の１週間前の授業時に１回目，最終発表の１週間前の授業時に２回目の評価者トレーニングを，下記の手順で行った。

1. 学生にルーブリックと評価シートを配布。
2. ルーブリックに書かれている評価項目および評価基準の説明を行う。
3. サンプル動画の中から一人（A 君）のビデオを見せて，評価させる。
4. 教員から A 君の問題点あるいは改善点についてコメントをする。
5. 評価シートを回収し，答え合わせを行う。

簡易グループのトレーニングに要した時間は約 15 分であった。

6. 分析 1: 評価者トレーニングは，学生の評価力に影響を及ぼすか？ また，どの評価項目に影響を及ぼすか？

6.1 評価者トレーニング１回目後の中間発表における学生による評価と教員による評価の相関

6.1.1 分析方法

学生の評価力を調べるため，教員による評価と同じように，優れた発表には高い評価を，あまりよくない発表には低い評価をしているかを調べた。学生が行ったプレゼンテーションに対する学生による相互評価と，教員による相互評価との相関係数を詳細グループ，簡易グループごとに算出した。学生一人ひとりのパフォーマンスに対する学生による評価の合計点の正規性の検定を行ったところ，表 3 が示す結果となった。

正規性が認められなかったことと，担当クラスの数の関係，さらに，5.2 で述べたように，両グループの等質性を確保するために，データ数が 24 名と少なくなったことにより，スピアマンの相関係数を用いた。次に，詳細グループ，簡易グループの相関係数に差があるかを調べるために，マン・ホイットニーの *U* 検定を行い，効果量を算出した。

表 3

正規性の検定結果

	Kolmogorov-Smimov(a)		
	統計量	自由度	有意確率
詳細グループ	.124	171	.000
簡易グループ	.068	171	.035

6.1.2 結果

6.1.2.1 グループ別の分析結果と考察

表 4 に，評価者トレーニングを行った詳細グループにおける，学生による相互評価と教員による評価との相関係数の記述統計量を示す。図 5 は，詳細グループの箱ひげ図である。

表 4

詳細グループの記述統計量　　（*n*=12）

	M	*SD*
内容	.301	.233
オリジナリティ	.149	.106
アイコンタクト	.493	.198
スライド	.170	.283
発音	.374	.328
ポーズ	.314	.258
合計	.383	.277

詳細グループにおいては，評価の合計点において，弱い相関が得られた（*r* = .383）。しかし，評価項目別に見ると，アイコンタクト（*r* = .493）においては，中程度の相関が得られたが，その他の項目，内容（*r* = .301），オリジナリティ（*r* = .149），スライド（*r* = .170），発音（*r* = .374），ポーズ（*r* = .314）においては，それぞれ弱い相関が得られた。

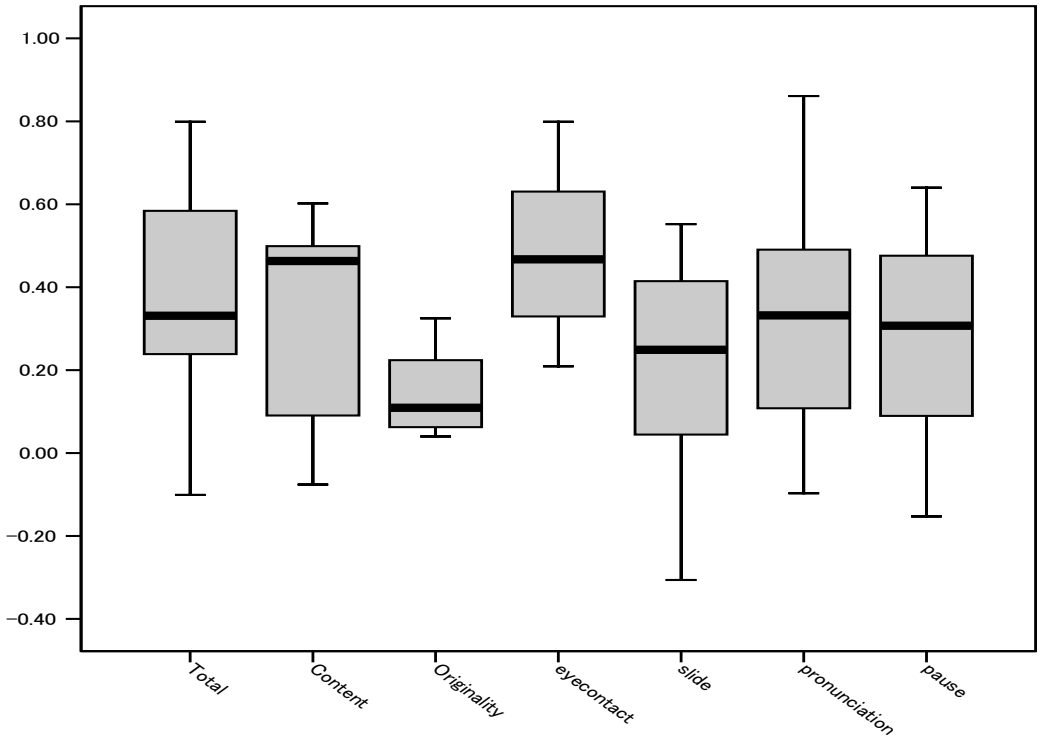


図 5. 評価者トレーニング 1 回目の詳細グループの箱ひげ図

図 5 からは，評価の合計（*SD* = .277）においては，強い相関（*r* = .799）からマイナ

スの関係（r = －.101）になるなど，相関係数が４段階に大きくばらついていることがわかった。また，評価項目別に見ると，「発音」（SD = .328）は，強い相関（r = .861）からマイナスの関係（r = －.097）になるなど，相関係数が４段階に大きくばらついており，さらに，「内容」（SD =.233）は，中程度（r = .602）からマイナスの関係（r = －.076），「スライド」（SD = .283）も，中程度（r = .552）からマイナスの関係（r = －.306），また，「ポーズ」（SD = .258）においても，中程度（r = .640）からマイナスの関係（r = －.153）になるなど，それぞれの相関係数が３段階に大きくばらついていることがわかった。

一方，「アイコンタクト」（SD = .198）は，強い相関（r = .799）から弱い相関（r = .209）とちらばりが小さく，教員による評価との相関も中程度であることから，この項目は，評価者トレーニングを行った学生にとって，その良し悪しがわかりやすく，評価しやすくなったものと考えられる。しかし，「オリジナリティ」（SD = .106）においては，標準偏差は小さいものの，教員による評価との相関が，弱い相関（r = .325）から相関がない程度（r = .040）に留まっていることから，この項目の学生の評価力は低く，「オリジナリティ」という評価項目は，たとえ評価者トレーニングを行ったとしても，学生にとって評価するのが難しかったものと考えられる。

これらの結果から，詳細グループの学生の評価と教員による評価の相関は高くないことがわかった。学生による評価と教員による評価の間に高い相関がないことは，学生の評価力は高くないと言える。つまり，詳細な評価者トレーニングを１回行った後の学生の評価力は高くないことがわかった。このことから，評価者トレーニングを１回行っただけでは，学生の評価力には影響しないことがわかった。

次に，簡易グループにおける，学生による相互評価と教員による評価との相関係数の記述統計量を表５に示す。図６は，簡易グループの箱ひげ図である。

表５

簡易グループの記述統計量　　（n=12）

	M	SD
内容	.076	.169
オリジナリティ	.341	.236
アイコンタクト	.611	.113
スライド	.157	.212
発音	.447	.230
ポーズ	.262	.373
合計	.383	.271

簡易グループにおいては，評価の合計点において，弱い相関が得られた（r = .383）。評価項目別に見ると，「内容」（r = .076），「オリジナリティ」（r = .341），「スライド」（r = .157），「ポーズ」（r = .262）においては，それぞれ弱い相関が得られたが，「アイコンタクト」（r = .611），「発音」（r = .477）においては，それぞれ中程度の相関が得られた。

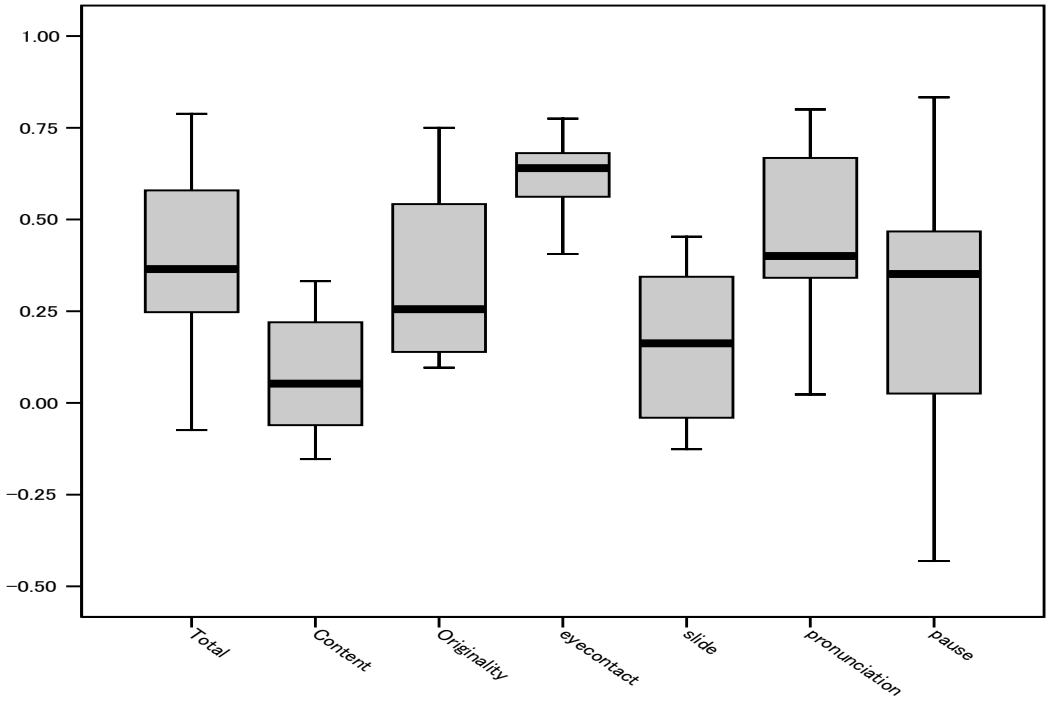


図 6. 簡易グループの箱ひげ図

また，図 6 からは，評価の合計（SD = .271）においては，高い相関（r = .788）からマイナスの関係（r = −.074）になるなど，相関係数が 4 段階に大きくばらついていることがわかった。また，評価項目別に見ると，内容」（SD = .169）においては，低い相関（r = .332）からマイナスの関係（r = −.153），「アイコンタクト」（SD = .113）においては，高い相関（r = .775）から中程度の相関（r = .406）とばらつきが小さいが，「オリジナリティ」（SD = .236）は，高い相関（r = .750）から相関がない関係（r = .096），「スライド」（SD = .212）は，中程度の相関（r = .453）からマイナスの関係（r = −.126），「発音」（SD = .230）は，高い相関（r = .800）から相関がない関係（r = .023）と 3 段階にまで大きくばらついている。さらに「ポーズ」（SD = .373）においては，高い相関（r = .833）からマイナスの関係（r = −.431）とさらに 4 段階にまで大きくばらついていることがわかった。

これらの結果から，簡易な評価者トレーニングを１回行ったグループの学生の評価と教員による評価の相関は高くないことがわかった。学生による評価と教員による評価の間に高い相関がないことは，学生の評価力は高くないと言える。つまり，簡易グループの評価力は低いことがわかった。このことから，簡易な評価者トレーニングを１回行っただけでは，学生の評価力には影響しないことがわかった。

6.1.2.2　評価項目別のグループ間の分析結果と考察

6.1.2.1 で評価者トレーニングの１回目を行った後の総合的な評価力は，詳細グループ，簡易グループのどちらも高くないことが確認された。そこで，詳細グループ，簡易グループそれぞれの学生の評価力が，評価項目によって違いがあるかを調べた。次に，詳細グループと簡易グループの相関係数に差があるかを調べるために，マン・ホイットニーの *U* 検定によって多重比較を行った。さらに効果量を算出した。

表 6 と表 7 に「内容」における教員による評価と学生による評価の相関係数の記述統計量と検定結果を示す。

表 6

評価項目「内容」の記述統計量

	n	*M*	*SD*	最大値	最小値
詳細グループ	12	.301	.233	.602	-.076
簡易グループ	12	.076	.169	.332	-.153

表 7

評価項目「内容」のマン・ホイットニーの *U* 検定結果・効果量

	U	*W*	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
詳細グループ VS 簡易グループ	32	110	-2.309	.020	-.472	中	詳細グループ

表 6 と表 7 が示すように，評価項目「内容」における教員による評価と学生による評価の相関係数は，詳細グループと簡易グループの間に差があり，効果量が中（$r = -.47$）であった。つまり，「内容」の項目においては，詳細グループの方が評価力が高いことがわかった。このことから，「内容」に対する評価力に，本格的な評価者トレーニングが影響を及ぼした可能性がある。

次に，表 8 と表 9 に「オリジナリティ」における教員による評価と学生による評価の相関係数の記述統計量と検定結果を示す。

表 8
評価項目「オリジナリティ」の記述統計量

	n	*M*	*SD*	最大値	最小値
詳細グループ	12	.149	.106	.325	.040
簡易グループ	12	.341	.236	.750	.096

表 9
評価項目「オリジナリティ」のマン・ホイットニーの *U* 検定結果・効果量

	U	*W*	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
詳細グループ VS 簡易グループ	30	108	-2.425	.014	-.496	中	簡易グループ

表 8 と表 9 が示すように，評価項目「オリジナリティ」における教員による評価と学生による評価の相関係数は，詳細グループと簡易グループの間に差があり，効果量は中（*r* = -.49）であった。つまり，「オリジナリティ」の項目においては，本格的な評価者トレーニングを行っていない簡易グループの評価力の方が高いことがわかった。このことから，評価者トレーニングの質や内容の違いは，トレーニングを 1 回行っただけでは，評価するのが難しいとされる「オリジナリティ」という評価項目に対する評価力には影響を及ぼさなかったと考えられる。また，詳細トレーニングがかえってこの項目に注意を向けることを邪魔した可能性も考えられる。

次に，表 10 と表 11 に「アイコンタクト」における教員による評価と学生による評価の相関係数の記述統計量と検定結果を示す。

表 10

評価項目「アイコンタクト」の記述統計量

	n	*M*	*SD*	最大値	最小値
詳細グループ	12	.493	.198	.799	.209
簡易グループ	12	.611	.113	.775	.406

表 11

評価項目「アイコンタクト」のマン・ホイットニーの *U* 検定結果・効果量

	U	*W*	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
詳細グループ VS 簡易グループ	47	125	-1.443	.160	-.295	小	簡易グループ

表 10 と表 11 が示すように，評価項目「アイコンタクト」における教員による評価と学生による評価の相関係数は, 詳細グループと簡易グループの間に有意差はないが, 効果量は小 (*r* = -.29) で差が見られた。つまり,「アイコンタクト」の項目においては, 本格的な評価者トレーニングを行っていない簡易グループの評価力の方が高いことがわかった。このことから,「アイコンタクト」は，その良し悪しが見た目にもはっきりわかりやすく，学生にとって評価しやすい評価項目ではあるが，１回だけの評価者トレーニングでは，その質や内容の違いは,「アイコンタクト」の対する評価力には影響を及ぼさなかったと考えられる。また，詳細トレーニングがかえってこの項目に注意を向けることを邪魔した可能性も考えられる。

次に，表 12 と表 13 に「スライド」における教員による評価と学生による評価の相関係数の記述統計量と検定結果を示す。

表 12

評価項目「スライド」の記述統計量

	n	*M*	*SD*	最大値	最小値
詳細グループ	12	.170	.283	.552	-.306
簡易グループ	12	.157	.212	.453	-.126

表 13

評価項目「スライド」のマン・ホイットニーの *U* 検定結果・効果量

	U	*W*	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
詳細グループ VS 簡易グループ	63	141	-.520	.630	-.107	小	詳細グループ

表 12 と表 13 が示すように，評価項目「スライド」における教員による評価と学生による評価の相関係数は，詳細グループと簡易グループの間に有意差はないが，効果量は小 (r = -.10) で差が見られた。つまり，「スライド」の項目においては，本格的なトレーニングを行った詳細グループの評価力が高いことがわかった。このことから，「スライド」はそのわかりやすさなどが見た目にわかりやすく評価しやすい項目と考えられるが，判定基準が書かれたルーブリックのみではなく，さらにスライドのサンプルを用いて本格的なトレーニングを行ったことで，その良し悪しがさらに明確になり，評価力に影響を及ぼしたと考えられる。

次に，表 14 と表 15 に「発音」における教員による評価と学生による評価の相関係数の記述統計量と検定結果を示す。

表 14

評価項目「発音」の記述統計量

	n	*M*	*SD*	最大値	最小値
詳細グループ	12	.374	.328	.861	-.097
簡易グループ	12	.447	.230	.800	.023

表 15

評価項目「発音」のマン・ホイットニーの *U* 検定結果・効果量

	U	*W*	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
詳細グループ VS 簡易グループ	63	141	-.520	.630	-.107	小	簡易グループ

表 14 と表 15 が示すように，評価項目「発音」における教員による評価と学生によ

る評価の相関係数は，詳細グループと簡易グループの間に有意差はないが，効果量が小 (r = -.10) で差が見られた。つまり，「発音」の項目においては，本格的な評価者トレーニングを行っていない簡易グループの評価力の方が高いことがわかった。このことから，「発音」は，目に見えず，聞こえてきた英語からその良し悪しを判断することになるため，印象によって左右されやすい評価項目であることと，また「発音」には，個々の発音の他にリズムやイントネーションなども含まれることから，それらのどこに注目するかで評価が変わってくることが考えられる。そのため，学生にとって，「発音」を評価させるのは難しいということも考えられる。このことから，ルーブリックとデモ音声を用いた本格的な評価者トレーニングを行った詳細グループの方が評価力が高くなると予想されたが，１回だけの評価者トレーニングでは，その質や内容の違いは，「発音」という評価項目には影響を及ぼさなかったと考えられる。

次に，表 16 と表 17 に「ポーズ」における教員による評価と学生による評価の相関係数の記述統計量と検定結果を示す。

表 16
評価項目「ポーズ」の記述統計量

	n	*M*	*SD*	最大値	最小値
詳細グループ	12	.314	.258	.640	-.153
簡易グループ	12	.262	.373	.833	-.431

表 17
評価項目「ポーズ」のマン・ホイットニーの U 検定結果・効果量

	U	*W*	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
詳細グループ							
VS	66	144	-.347	.755	-.071	なし	
簡易グループ							

表 16 と表 17 が示すように，評価項目「ポーズ」における教員による評価と学生による評価の相関係数は，詳細グループと簡易グループの間に差はなかった。つまり，「ポーズ」の項目においては，評価力の差がないことがわかった。「ポーズ」も「発音」同様，聞こえてきた英語で評価せざるを得ない項目である。このことから，例文を用い，適切なポーズの位置を例示した本格的なトレーニングを行った詳細グループの評価力の方が高くなると予想された。しかし，チャンキングの能力というのは，正しく

意味を理解できているかということに影響されており，評価者トレーニングを行っただけではそのスキルは伸びない。そのため，評価者トレーニングの影響を及ぼしにくい項目であると考えられる。

上述の分析結果を表 18 にまとめた。

表 18
分析結果のまとめ

	内容	オリジナリティ	アイコンタクト	スライド	発音	ポーズ	合計
詳細グループ VS 簡易グループ	詳細グループ	簡易グループ	簡易グループ	詳細グループ	詳細グループ		

評価者トレーニングは，評価力を高めるために行われるものであることから，評価者トレーニングを行った後の学生の評価力は，本格的に評価者トレーニングを行った詳細グループの評価力が，簡易グループの評価力より高くなるものと予想していた。しかし，両グループの評価力を等質にしたため，合計では差が出なかったが，表18が示すように，評価項目別には差が見られた。つまり，１度の評価者トレーニングの質や内容の違いでは，学生の評価力に影響を及ぼすとは言えない。そこで，評価者トレーニングの2回目を実施し，トレーニングの１回目と２回目では学生の評価力に差が出るかを調べることにより，評価者トレーニングの回数が学生の評価力に影響を及ぼすかを検証した。

7. 分析 2：評価者トレーニングの回数は，学生の評価力に影響を及ぼすか？ また，どの評価項目に影響を及ぼすか？

7.1 評価者トレーニング２回目後の最終発表における学生による評価と教員による評価の相関

7.1.1 分析方法

最終発表の１週間前の授業時に，評価者トレーニング２回目を実施し，最終発表時の詳細グループ，簡易グループそれぞれの学生による評価と教員による評価の相関を調べた。検定の方法は分析１と同じである。

7.1.2　結果

7.1.2.1　グループ別の分析結果と考察

表 19 に，評価者トレーニングを行った詳細グループにおける，学生による相互評価と教員による評価との相関係数の記述統計量を示す。

表 19

詳細グループの記述統計量　（*n*=12）

	M	*SD*
内容	.309	.222
オリジナリティ	.305	.259
アイコンタクト	.645	.094
スライド	.297	.278
発音	.615	.202
ポーズ	.596	.177
合計	.619	.181

詳細グループにおいては，評価の合計点において，中程度の相関が得られた（*r* = .619）。

しかし，評価項目別に見てみると，アイコンタクト（*r* = .645），発音（*r* = .615），ポーズ（*r* = .596）においてはそれぞれ中程度の相関が得られたが，内容（*r* = .309），オリジナリティ（*r* = .305），スライド（*r* = .297）においては，弱い相関が得られた。

次に，詳細グループの箱ひげ図を図 7 に示す。

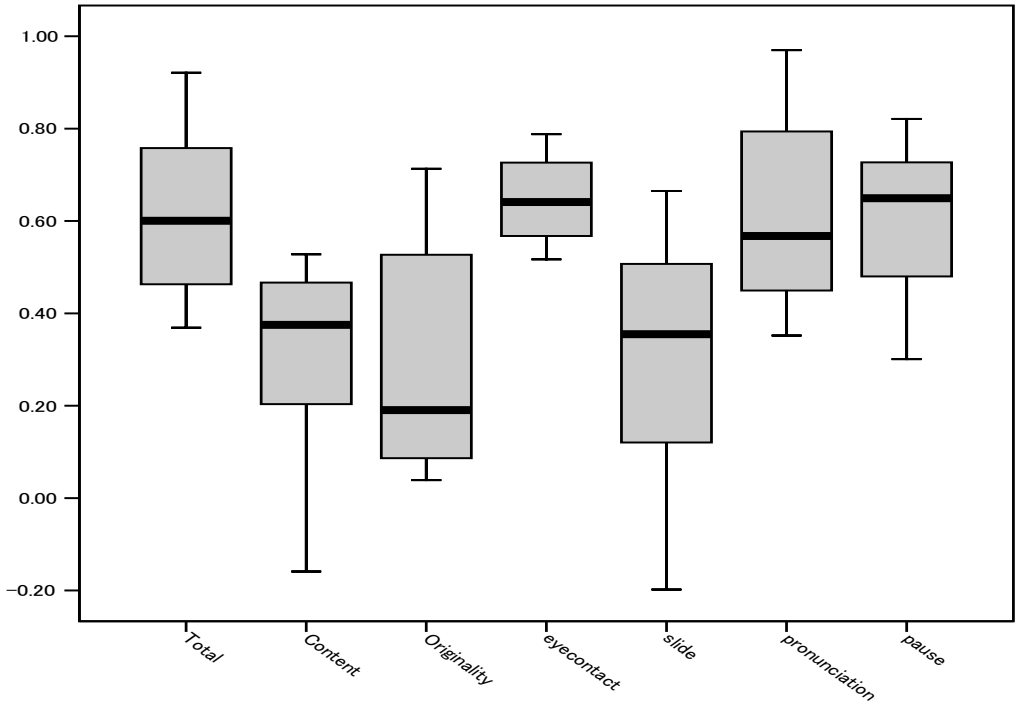


図 7. 詳細グループの箱ひげ図

図 7 からは，評価の「合計」（*SD* = .181）においては，高い相関（*r* = .921）から低い相関（*r* = .369）になり，１回目の評価者トレーニング後より，評価のばらつきが小さくなっていることがわかった。評価項目別に見ると，「アイコンタクト」（*SD* = .094）においては，高い相関（*r* = .788）から中程度の相関（*r* = .517）になり，相関係数のばらつきはあまり見られないことがわかった。しかし，「発音」（*SD* = .202）においては，高い相関（*r* = .970）から低い相関（*r* = .352）「ポーズ」（*SD* = .177）においては，（*r* = .821）から低い相関（*r* = .301）になるなど，それぞれの相関係数が２段階にばらついていることがわかった。

また，「内容」（*SD* = .222）においては，中程度の相関（*r* = .528）からマイナスの関係（*r* = −.159），「オリジナリティ」（*SD* = .259）においては，高い相関（*r* = .713）から相関がない関係（*r* = .039），「スライド」（*SD* = .278）は，中程度の相関（*r* = .665）からマイナスの関係（*r* = −.198）と，それぞれの相関係数が３段階に大きくばらついていることがわかった。これらの結果から，詳細グループにおいては，２回の詳細な評価者トレーニングは学生の評価力に影響を与えた可能性がある。しかし，「内容」，「オリジナリティ」，および「スライド」の評価項目に対する評価力は，評価者トレーニングを２回行った後でも，学生にとって評価するのに難しい項目であると考えられる。

次に，表 20 に，簡易グループにおける，学生による相互評価と教員による評価との相関係数の記述統計量を示す。

表 20
簡易グループの記述統計量　（*n*=12）

	M	*SD*
内容	.430	.279
オリジナリティ	.277	.250
アイコンタクト	.601	.214
スライド	.289	.377
発音	.446	.247
ポーズ	.170	.443
合計	.625	.131

表 20 が示すように，簡易グループにおいても，評価の合計点において，中程度の相関が得られた（*r* = .625）。しかし，評価項目別に見てみると，アイコンタクト（*r* = .601），発音（*r* = .446）内容（*r* = .430）においては，中程度の相関が得られたが，オリジナリティ（*r* = .277），スライド（*r* = .289），ポーズ（*r* = .170）においては，弱い相関が得られた。

図 8 に簡易グループの箱ひげ図を示す。

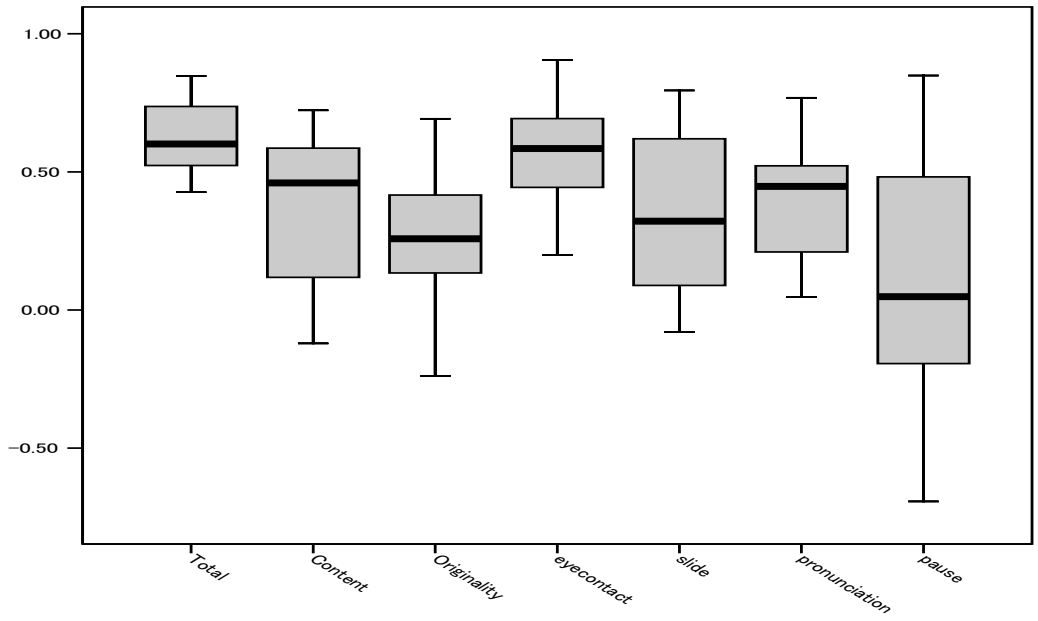


図 8. 簡易グループの箱ひげ図

図 8 からは，評価の「合計」（*SD* = .131）においては，高い相関（*r* = .847）から低い相関（*r* = .427）になり，詳細グループと同じように，１回目の評価者トレーニング後より，評価のばらつきが小さくなっていることがわかった。しかし，評価項目別に見てみると，「オリジナリティ」（*SD* = .250）においては，中程度の相関（*r* = .692）からマイナスの関係（*r* = −.239），「アイコンタクト」（*SD* = .214）においては，高い相関（*r* = .925）から相関がない関係（*r* = .199），「発音」（*SD* = .247）においても，高い相関（*r* = .768）から相関がない関係（*r* = .047）になるなど，それぞれ相関の程度が 3 段階に大きくばらついていることがわかった。

さらに，「内容」（*SD* = .279）においては，高い相関（*r* = .723）からマイナスの関係（*r* = −.121），「ポーズ」（*SD* = .443）においても，高い相関（*r* = .849）からマイナスの関係（*r* = −.693）となり，また「スライド（*SD* = .377）も，高い相関（*r* = .795）からマイナスの関係（*r* = −.504）になるなど，それぞれ相関の程度が 4 段階に大きくばらついていることがわかった。

これらの結果から，簡易グループにおいては，１回目の簡易な評価者トレーニング後の評価力と比べると，評価の「合計」は高くなったが，評価項目別の評価力は決して高くなったとは言えない。つまり，簡易なトレーニングの場合，たとえトレーニングを２回行っても，学生の評価力に影響を与えないと言えよう。

7.1.2.2　評価者トレーニング 1 回目と 2 回目における評価項目別のグループ間の分析結果と考察

7.1.2.1 で評価者トレーニングの 2 回目を行った後の全体的な評価力は，詳細グループ，簡易グループのどちらも高くなったことが確認された。次に，詳細グループと簡易グループの相関係数に差があるかを調べるために，マン・ホイットニーの *U* 検定によって多重比較を行った。さらに効果量を算出した。

そして，各グループにおける評価力の伸びを調べるために，評価者トレーニングの１回目を行った後と，２回目を行った後の学生の評価力の差を，ウィルコクソンの符号付順位検定を用いて調べた。

表 21 と表 22 に「合計」における教員による評価と学生による評価の相関係数の記述統計量と検定結果を示す。

表 21

評価の「合計」の記述統計量

	n	*M*	*SD*	最大値	最小値
詳細グループ	12	.619	.181	.921	.369
簡易グループ	12	.625	.131	.847	.427

表 22

評価の「合計」のマン・ホイットニーの *U* 検定結果・効果量

	U	*W*	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
詳細グループ VS 簡易グループ	70.5	148.5	-.087	.932	-.018	なし	

表 21 と表 22 から，教員による評価と学生による評価の相関係数は，評価の合計において，詳細グループと簡易グループの間に差はないことがわかる。つまり，全体的な評価力において差はない。

次に，表 23 に詳細グループの「合計」における評価トレーニング１回目後と２回目有意後の学生の評価力の差の検定結果および効果量を示す。

表 23

詳細グループの「合計」のウィルコクソンの符号付き順位検定結果および効果量

評価項目		トレーニング１回目	トレーニング２回目	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
合計	*M*	.383	.619	-2.67	.008	-.77	大	２回目
	SD	.277	.181					

表 23 から，詳細グループの評価の「合計」においては，トレーニングの１回目後の評価力と２回目後の評価力の間に差があり，効果量は大で，１回目後より２回目後の学生の評価力は伸びていることがわかる。つまり，本格的な評価者トレーニングを２回行うことで学生の評価力を高めることができると言えよう。

次に，表 24 に簡易グループの「合計」における評価者トレーニング１回目と２回目の学生の評価力の差の検定結果および効果量を示す。

表 24

簡易グループの「合計」のウィルコクソンの符号付き順位検定結果および効果量

評価項目		トレーニング１回目	トレーニング２回目	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
合計	*M*	.383	.625	-2.59	.01	-.748	大	２回目
	SD	.271	.131					

表 24 から，簡易グループの評価の「合計」においては，トレーニングの１回目後の評価力と２回目後の評価力の間に差があり，効果量は大で，簡易グループにおいても，１回目後より２回目後の学生の評価力は有意に伸びていることがわかる。つまり，簡易なトレーニングであっても，２回実施することで，学生の評価力を高められる可能性があることがわかった。

表 23 と表 24 から，本格的な評価者トレーニング，簡易なトレーニングをそれぞれ２回行っても，両グループの全体的な評価力に差はでなかったことから，評価者トレーニングの内容や質よりも，２回というトレーニングの実施回数の方が学生の評価力に影響を及ぼした可能性がある。

次に，評価項目別にグループ間の分析を行った。

表 25 と表 26 に「内容」における教員による評価と学生による評価の相関係数の記述統計量と検定結果を示す。

表 25

評価項目「内容」の記述統計量

	n	*M*	*SD*	最大値	最小値
詳細グループ	12	.309	.222	.528	-.159
簡易グループ	12	.430	.279	.723	-.121

表 26

評価項目「内容」のマン・ホイットニーの *U* 検定結果・効果量

	U	*W*	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
詳細グループ VS 簡易グループ	44	122	-1.617	.114	-.33	中	簡易グループ

表 25 と表 26 から，評価項目「内容」における教員による評価と学生による評価の相関係数は，詳細グループと簡易グループの間に有意差はないが，効果量は中 (*r* = -.33) で差が見られた。つまり，「内容」の項目においては，本格的な評価者トレーニングを行っていない簡易グループの評価力の方が高いことがわかった。

表 27 に詳細グループの「内容」における評価トレーニング１回目後と２回目後の学生の評価力の差の検定結果および効果量を示す。

表 27

詳細グループの「内容」のウィルコクソンの符号付き順位検定結果および効果量

評価項目		トレーニング１回目	トレーニング２回目	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
内容	*M*	.301	.309	-.235	.814	-.068	なし	
	SD	.233	.222					

表 27 から，詳細グループの「内容」におけるトレーニング１回目後と２回目後の間の評価力に差はないことがわかる。つまり，本格的な評価者トレーニングを２回行っても，「内容」という項目においては，学生の評価力を上げることができないということである。

次に，表 28 に簡易グループの「内容」における評価トレーニング１回目後と２回目後の学生の評価力の差の検定結果および効果量を示す。

表 28

簡易グループの「内容」のウィルコクソンの符号付き順位検定結果および効果量

評価項目		トレーニング１回目	トレーニング２回目	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
内容	*M*	.076	.430	-2.67	.01	-.77	大	２回目
	SD	.169	.279					

表 28 から，簡易グループの「内容」においては，トレーニング１回目後と２回目後の評価力の間に差があり，効果量は大で，１回目後より２回目後の学生の評価力は有意に伸びていることがわかる。つまり，簡易な評価者トレーニングを２回行うことで，「内容」という項目対する学生の評価力に影響を及ぼした可能性がある。

表 27 と表 28 から，１回目の評価者トレーニング後と２回目後とでは，詳細グループと簡易グループの評価力が逆転していることがわかる。これは,「内容」の評価には，評価者トレーニングより，他の要因，例えば内容を理解するために必要な英語力や理解の助けとなるであろう予備知識の有無などが，評価力に影響を及ぼしている可能性があると考えられる。

次に，表 29 と表 30 に「オリジナリティ」における教員による評価と学生による評価の相関係数の記述統計量と検定結果を示す。

表 29

評価項目「オリジナリティ」の記述統計量

	n	*M*	*SD*	最大値	最小値
詳細グループ	12	.305	.259	.713	.039
簡易グループ	12	.277	.250	.692	-.239

表 30

評価項目「オリジナリティ」のマン・ホイットニーの *U* 検定結果・効果量

	U	*W*	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
詳細グループ							
VS	65	143	-.062	.976	-.013	なし	
簡易グループ							

表 29 と表 30 から，評価項目「オリジナリティ」における教員による評価と学生による評価の相関係数は，詳細グループと簡易グループの間に差はないことがわかる。つまり,「オリジナリティ」に対する評価力において差はない。

表 31 に詳細グループの「オリジナリティ」における評価者トレーニング１回目後と２回目後の学生の評価力の差の検定結果および効果量を示す。

表 31

詳細グループの「オリジナリティ」のウィルコクソンの符号付き順位検定結果および効果量

評価項目		トレーニング１回目	トレーニング２回目	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
オリジナリティ	*M*	.149	.305	-1.689	.091	-.51	大	２回目
	SD	.106	.259					

表 31 から，詳細グループの「オリジナリティ」においては，トレーニング１回目後と２回目後の間の評価力に有意差はないが，効果量は大 (r = -.51) で差が見られ，「オリジナリティ」という項目における詳細グループの評価力は有意に伸びていることがわかる。つまり，本格的な評価者トレーニングを２回行うことで，この項目に対する評価力を高められる可能性があると言えよう。

次に，表 32 に簡易グループの「オリジナリティ」における評価者トレーニング１回目後と２回目後の学生の評価力の差の検定結果および効果量を示す。

表 32

簡易グループの「オリジナリティ」のウィルコクソンの符号付き順位検定結果および効果量

評価項目		トレーニング１回目	トレーニング２回目	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
オリジナリティ	*M*	.341	.277	-.58	.56	-.175	小	１回目
	SD	.236	.250					

表 32 から，簡易グループの「オリジナリティ」においては，トレーニング１回目後と２回目後の間の評価力に有意差はないが，効果量は小 (r = -.17) で差が見られ，トレーニング２回目後より１回目後の評価力の方が高いことがわかる。つまり，簡易グループにおいては，トレーニングを２回行った後の学生の評価力が逆に下がっていることがわかった。

表 31 と表 32 から，「オリジナリティ」という項目においては，トレーニングの回数より，トレーニングの質や内容や２つのグループでおこなわれたプレゼンの違いが，学生の評価力に影響を及ぼした可能性があると言えよう。「オリジナリティ」という評

価項目には「内容」と同じく，発表内容を理解する英語力やその内容に対する予備知識の有無などが，学生の評価力に影響を及ぼしている可能性があると考えられる。しかし，詳細なトレーニングを２回行うことによって，主観的になりがちと考えられる「オリジナリティ」に対する評価が，ルーブリックに記載されている判定基準に基づいた，より客観的に評価できるようになった可能性を示している。

次に，表 33 と表 34 に「アイコンタクト」における教員による評価と学生による評価の相関係数の記述統計量と検定結果を示す。

表 33
評価項目「アイコンタクト」の記述統計量

	n	*M*	*SD*	最大値	最小値
詳細グループ	12	.645	.094	.788	.517
簡易グループ	11	.601	.214	.925	.199

表 34
評価項目「アイコンタクト」のマン・ホイットニーの *U* 検定結果・効果量

	U	*W*	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
詳細グループ VS 簡易グループ	55	121	-.677	.525	-.139	小	詳細グループ

表 33 と表 34 から，評価項目「アイコンタクト」における教員による評価と学生による評価の相関係数は，詳細グループと簡易グループの間に有意差はないが，効果量は小 (r = -.13) で差が見られたことから，「アイコンタクト」の項目においては，詳細グループの評価力が高いことがわかる。つまり，「アイコンタクト」の項目においては，トレーニングの回数より，質や内容の違いが学生の評価力に影響を及ぼした可能性があると言えよう。

表 35 に詳細グループの「アイコンタクト」における評価者トレーニング１回目後と２回目後の学生の評価力の差の検定結果および効果量を示す。

表 35

詳細グループの「アイコンタクト」のウィルコクソンの符号付き順位検定結果および効果量

評価項目		トレーニング１回目	トレーニング２回目	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
アイコンタクト	*M*	.493	.645	-2.040	.041	-.589	大	２回目
	SD	.198	.094					

表 35 から，詳細グループの「アイコンタクト」における，トレーニング１回目後と２回目後の評価力の間には差があり，効果量は大（r = -.58）で,「アイコンタクト」における詳細グループの評価力は有意に伸びていることがわかる。詳細グループは，１回目の評価トレーニングでは筆者の実演を見ただけで評価を行わなくてはならなかったが, 評価トレーニング２回目の時では, クラスメイトの発表の評価経験だけでなく, 学生自身の発表経験も積んだことで,「アイコンタクト」をどのように取るのがより説得力のある良いプレゼンテーションとなるのかについてさらに理解を深めたことが, 評価力の向上につながったと考えられる。

次に，表 36 に簡易グループの「アイコンタクト」における評価者トレーニング１回目後と２回目後の学生の評価力の差の検定結果および効果量を示す。

表 36

簡易グループの「アイコンタクト」のウィルコクソンの符号付き順位検定結果および効果量

評価項目		トレーニング１回目	トレーニング２回目	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
アイコンタクト	*M*	.611	.601	.00	1.00	0	なし	
	SD	.113	.214					

表 36 から，簡易グループの「アイコンタクト」における，トレーニング１回目後と２回目後の評価力の間に差がないことがわかる。つまり，簡易なトレーニングを２回行っても，学生の「アイコンタクト」に対する学生の評価力を上げることはできないこと考えられる。

また，表 35 と表 36 が示すように,「アイコンタクト」に対する評価力は，トレーニ

ングの実施回数よりも，トレーニング内容やその質が，影響を及ぼすと考えられる。

次に，表 37 と表 38 に「スライド」における教員による評価と学生による評価の相関係数の記述統計量と検定結果を示す。

表 37

評価項目「スライド」の記述統計量

	n	*M*	*SD*	最大値	最小値
詳細グループ	12	.297	.278	.665	-.198
簡易グループ	12	.289	.377	.795	-.504

表 38

評価項目「スライド」のマン・ホイットニーの *U* 検定結果・効果量

	U	*W*	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
詳細グループ							
VS	70	148	-.115	.932	-.024	なし	
簡易グループ							

表 37 と表 38 から，評価項目の「スライド」において，詳細グループと簡易グループの間に差がないことがわかる。つまり，「スライド」に対する評価力において差はない。このことは，「スライド」に」対する評価力は，評価者トレーニングの質や内容には影響されないと考えられる。

表 39 に詳細グループの「スライド」における評価者トレーニング１回目後と２回目後の学生の評価力の差の検定結果および効果量を示す。

表 39

詳細グループの「スライド」のウィルコクソンの符号付き順位検定結果および効果量

評価項目		トレーニング１回目	トレーニング２回目	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
スライド	*M*	.170	.297	-1.334	.182	-.385	中	２回目
	SD	.283	.278					

表 39 から，詳細グループの「スライド」における，トレーニング１回目後と２回目

後の評価力の間に差があり，効果量は中（r = -.38）であることがわかる。このことから，本格的なトレーニングを２回行うことで,「スライド」に対する学生の評価力を向上させることができたと言える。「スライド」というのは，その良し悪しが見た目にはっきりとわかりやすいため，良いスライドとあまり良くないスライドを何度か目にすることで，スライドの良し悪しがわかるようになると考えられる。つまり，評価者トレーニングの内容だけではなく，その回数が「スライド」に対する評価力に影響を及ぼすと言えよう。

次に，表 40 に簡易グループの「スライド」における評価者トレーニング１回目後と２回目後の学生の評価力の差の検定結果および効果量を示す。

表 40

簡易グループの「スライド」のウィルコクソンの符号付き順位検定結果および効果量

評価項目		トレーニング１回目	トレーニング２回目	Z	p	効果量 r	効果の大きさ	優劣関係
スライド	*M*	.157	.289	-1.10	.27	-.318	中	２回目
	SD	.212	.377					

表 40 から，簡易グループにおいても，トレーニング１回目後と２回目後の「スライド」における評価力に有意差はないが，効果量は中 (r = -.31) で差が見られた。このことから，簡易なトレーニングであっても，２回実施することで,「スライド」という項目に対する学生の評価力を向上させることができることがわかった。

表 39 と表 40 から,「スライド」は，その良し悪しが見た目に分かりやすく評価しやすい項目であり，また発表の内容が反映されたものであることから，評価者トレーニングの回数を重ね，判定基準をしっかりと理解できれば評価力はついてくると考えられる。そのため，１回目後は詳細グループの評価力の方が高かったが，評価者トレーニングを２回行ったことで，簡易グループの学生にもこの項目に対する評価力がついてきたと考えられる。つまり,「スライド」の項目には，トレーニングの内容や質よりも，回数の方が学生の評価力に影響を及ぼしたと考えられる。

次に，表 41 と表 42 に「発音」における教員による評価と学生による評価の相関係数の記述統計量と検定結果を示す。

表 41

評価項目「発音」の記述統計量

	n	*M*	*SD*	最大値	最小値
詳細グループ	12	.615	.202	.97	.352
簡易グループ	12	.446	.247	.768	.047

表 42

評価項目「発音」のマン・ホイットニーの *U* 検定結果・効果量

	U	*W*	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
詳細グループ VS 簡易グループ	46	124	-1.502	.143	-.307	中	詳細グループ

表 41 と表 42 から，評価項目「発音」における教員による評価と学生による評価の相関係数は，詳細グループと簡易グループの間に有意差はないが，効果量は中 (*r* = -.30) で差が見られた。つまり，「発音」の項目においては，詳細グループの評価力が高いことがわかった。このことは，簡易グループの評価者トレーニングでは，「発音」の項目についての判定基準の説明のみ行ったのに対し，詳細グループに評価者トレーニングでは，例文と音声のデモテープを用いて，より具体的に各レベルの説明をしたうえでトレーニングを行ったことが，学生の評価力に影響を与えたのであろう。

表 43 に詳細グループの「発音」における評価者トレーニング１回目後と２回目後の学生の評価力の差の検定結果および効果量を示す。

表 43

詳細グループの「発音」のウィルコクソンの符号付き順位検定結果および効果量

評価項目		トレーニング１回目	トレーニング２回目	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
発音	*M*	.374	.615	-1.961	.050	-.567	大	２回目
	SD	.328	.202					

表 43 から，詳細グループの「発音」における，評価者トレーニングの１回目後と２回目後の学生の評価力には差があり，効果量は大（*r* = -.56）であることがわかる。こ

のことは，上述の本格的なトレーニングを２回行ったことで,「発音」という項目に対する学生の評価力を大幅に高めたと言えよう。

次に，表 44 に簡易グループの「発音」における評価者トレーニング１回目後と２回目後の学生の評価力の差の検定結果および効果量を示す。

表 44

簡易グループの「発音」のウィルコクソンの符号付き順位検定結果および効果量

評価項目		トレーニング１回目	トレーニング２回目	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
発音	*M*	.447	.446	-.08	.94	-.023	なし	
	SD	.230	.247					

表 44 から，簡易グループにおいては，評価者トレーニングの１回目後と２回目後の学生の評価力に差はないことがわかる。上述の簡易なトレーニングの場合，たとえ２回トレーニングを実施しても,「発音」という項目の評価力を上げることができないと言えよう。

表 43 と表 44 から，評価者トレーニングの１回目後の評価も，２回目後の評価においても，本格的な評価者トレーニングを行った詳細グループの評価力が簡易グループの評価力を上回り，またその差も大きくなったことから，評価者トレーニングの回数より質が「発音」対する評価力に影響を及ぼすと思われる。

次に，表 45 と表 46 に「ポーズ」における教員による評価と学生による評価の相関係数の記述統計量と検定結果を示す。

表 45

評価項目「ポーズ」の記述統計量

	n	*M*	*SD*	最大値	最小値
詳細グループ	12	.596	.177	.821	.301
簡易グループ	12	.170	.443	.849	-.693

表 46

評価項目「ポーズ」のマン・ホイットニーの *U* 検定結果・効果量

	U	*W*	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
詳細グループ VS 簡易グループ	30	108	-2.425	.014	-.495	中	詳細グループ

表 45 と表 46 から，評価項目「ポーズ」における教員による評価と学生による評価の相関係数は，詳細グループと簡易グループの間に差があり，効果量は中（*r* = -.49）であることがわかる。つまり，「ポーズ」の項目においては，詳細グループの評価力が高いことがわかった。これは，「ポーズ」に対する評価力には，評価者が正しく発話をチャンキングできていることと正しく意味理解ができていることが，評価者トレーニングより影響を及ぼすと考えていたが，本格的なトレーニングを２回行うことにより，クラスメイトの発表を意味のまとまり毎に聞いて内容を理解しようする意識が芽生え，評価力の向上につながっていったと考えられる。

表 47 に詳細グループの「ポーズ」の検定結果および効果量を示す。

表 47

詳細グループの「ポーズ」のウィルコクソンの符号付き順位検定結果および効果量

評価項目		トレーニング１回目	トレーニング２回目	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
ポーズ	*M*	.314	.596	-2.432	.015	-.703	大	２回目
	SD	.258	.177					

表 47 から，詳細グループの「ポーズ」における，評価者トレーニング１回目後と２回目後の学生の評価力に差があり，効果量は大であることがわかる。このことから，「ポーズ」という項目における詳細グループの評価力は有意に伸びていると考えられる。つまり，本格的な評価者トレーニングを２回行うことで，この項目に対する評価力を高めることができると言えよう。

次に，表 48 に簡易グループの「ポーズ」の検定結果および効果量を示す。

表 48

簡易グループの「ポーズ」のウィルコクソンの符号付き順位検定結果および効果量

評価項目		トレーニング１回目	トレーニング２回目	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
ポーズ	*M*	.262	.170	-.71	.48	-.204	小	１回目
	SD	.373	.443					

表 48 から，簡易グループにおいては，トレーニング１回目後と２回目後の間の評価力に有意差はないが，効果量は小 (r = -.20) で差が見られ，トレーニング２回目後より１回目後の評価力の方が高いことがわかる。つまり，簡易グループにおいては，トレーニングを２回行った後の学生の評価力が逆に下がっていることがわかった。

１回目の評価者トレーニング後の評価力は，両グループ間に差はなく，２回目の評価者トレーニング後の評価力に差があるということは，トレーニング中に発話の具体例を示した評価者トレーニングがチャンキングや意味理解に注意を向けながら評価を行うことを促進したと考えられよう。

上述の分析結果を表 49 にまとめた。

表 49

分析結果のまとめ

	内容	オリジナリティ	アイコンタクト	スライド	発音	ポーズ	合計
詳細グループ VS 簡易グループ	簡易グループ		詳細グループ		詳細グループ	詳細グループ	

評価者トレーニングの２回目を行った後の最終発表時の学生の評価力は，評価の合計において,両グループの間に差はなかった。しかし，評価項目別に見ると，「アイコンタクト」，「発音」，「ポーズ」において差が見られ，詳細グループの評価力の方が高いことが確認された。このことから, 本格的な評価者トレーニングを２回行うことで, 上述の３つの項目に対する評価力を上げることができると言えよう。しかし，本来ならば，評価者トレーニングは評価力を高めるために行われるものであることから，本格的な評価者トレーニングを行った詳細グループの評価力が，すべての評価項目にお

いて高くなると予想されたが，表 49 が示すように，そのようにはならなかった。予想とは反対に，「内容」においては，本格的な評価者トレーニングを２回行った詳細グループより，簡易的なトレーニングのみ２回行った簡易グループの評価力の方が上回る結果となった。これは，「内容」の評価に対する評価力は，評価者トレーニングの質よりも，発表内容に対する予備知識や，内容を十分理解できるだけの英語力といった他の要因が関係していると考えられる。

また，簡易な評価者トレーニングを２回行った簡易グループの評価力は，全体的な評価力を上げることはできたが，「内容」を除くすべての評価項目に対する評価力を上げることができなかった。この結果から，簡易な評価者トレーニングの場合，回数を増やしたことで，学生の評価経験の浅さは改善できたかもしれないが，学生の評価力には影響を及ぼさないことがわかった。

また，評価者トレーニング１回目後から２回目後への評価力の伸びについては，詳細グループと簡易グループでは，「ポーズ」「内容」「オリジナリティ」の項目において，評価力の向上に違いが見られた。「ポーズ」を感知する能力は正しくチャンキングできることが前提となり，意味理解に影響を与える項目である。一方，「内容」，「オリジナリティ」の項目は，発表内容をどれだけ理解できているかということに大きく関係している。いずれも評価者トレーニングが影響を及ぼしにくい項目であったためと考えられる。

8. アンケート調査

8.1 材料

評価者トレーニングに対する学生の反応を調べるために，質問紙に選択肢式の項目を４つ作成し，対象者に評価最終日の授業時にアンケート調査を行った。

表 50 に選択肢式の項目内容と選択肢を示す。

表 50

アンケート調査の項目内容および選択肢

項目内容	選択肢
Q1.プレゼンテーションの評価について	教員のみによる評価が良い 学生のみによる評価が良い 教員，学生の両方による評価が良い
Q2.「評価項目、評価尺度一覧表（ルーブリック）」のわかりやすさについて	とてもわかりやすい わかりやすい どちらとも言えない わかりにくい とてもわかりにくい
Q3[1].評価トレーニング」は実際の評価をする際、役に立ちましたか。	非常に役に立った 役に立った どちらとも言えない 役に立たなかった 全く役に立たなかった
Q4.評価者としての経験は、あなたにとってどのような影響を与えましたか。	自由記述回答
Q5.学生が評価することのメリット、デメリットは何ですか。	自由記述回答

8.2 結果と考察

表 51 に，質問 1 の「プレゼンテーションの評価について」の回答結果を示す。

表 51

質問 1．「プレゼンテーションの評価について」の回答結果

（人数）

	教員のみがよい	学生のみがよい	教員と学生の両方が良い	計
詳細グループ	0	0	12	12
簡易グループ	0	1	11	12
計	0	1	23	24

表 52

ｶｲ 2 乗検定結果

	値	自由度	漸近有意確率 (両側)	効果量	効果の大きさ
Pearson のｶｲ 2 乗	1.043	1	.307	.209	小
連続修正(a)	0	1	1		
尤度比	1.429	1	.231		
Fisher の直接法					
線型と線型による連関	1	1	.317		
有効なｹｰｽの数	24				

χ^2検定を行ったところ，統計的な有意差は認められず，詳細グループと簡易グループとの間にはプレゼンテーションの評価についての反応に関連がなかった（表 52)。両グループとも，プレゼンテーションの評価をするのは教員のみが良いと回答した学生はおらず，評価は教員と学生の両方で行うのが良いと回答した。学生の回答結果から，学生が教員のみによる評価を望んでおらず，学生による評価も加えて欲しいと考えていることがわかる。その理由として，学生が教員のみによる評価には自分が正当に評価されていないと感じている可能性が考えられる。

表 53 に，質問 2. ルーブリックについての回答結果を示す。

表 53

質問 2.「ルーブリックについて」の回答結果

（件数）

	とても わかりやすい	わかりやすい	どちらとも 言えない	わかりにくい	とても わかりにくい	計
詳細グループ	6	5	1	0	0	12
簡易グループ	6	6	0	0	0	12
合計	12	11	1	0	0	24

表 54

ｶｲ 2 乗検定結果

	値	自由度	漸近有意確率 (両側)	効果量	効果の大きさ
Pearson のｶｲ 2 乗	1.090	2	.579	.209	小
尤度比	1.477	2	.477		
線型と線型による連関	0.120	1	.728		
有効なｹｰｽの数	24				

χ^2 検定を行ったところ，統計的な有意差は認められず，詳細グループと簡易グループとの間にはルーブリックのわかりやすさについての反応に関連がなかった（表 54）。ほぼすべての学生から，評価者トレーニングおよび実際の評価に用いたルーブリックはわかりやすいとの回答が得られた。筆者が勤務していたプログラムでは，評価項目と 5 段階の数字のみの判定基準が書かれた評価シートを用いて相互評価を行っているため，判定基準が明確でなく，評価者によってぶれてしまう可能性があった。しかし，評価項目に対する判定基準が明確に書かれたルーブリックは学生にとってわかりやすく，また評価しやすいものであったと考えられる。このことは，学生に相互評価をさせる際には，上述のようなルーブリックを用いて学生に相互評価をさせるべきであると考えられる。

表 55 に,質問 3.「評価者トレーニングは実際の評価をする際､役に立ちましたか（詳細グループのみ)」の回答結果を示す。

表 55

質問 3.「評価者トレーニングは実際の評価をする際、役に立ちましたか（詳細グループのみ)」の回答結果

（件数）

	非常に役に立った	役に立った	どちらとも言えない	役に立たなかった	全く役に立たなかった	計
詳細グループ	8	4	0	0	0	12

これまで，評価項目と判定基準のみ書かれた評価シートを用いて評価を行うことに慣れている学生にとっては，評価者トレーニングは初めての経験ということもあり，もしかすると中にはかえって困惑する学生もいるかもしれないと懸念された。しかし，本格的な評価者トレーニングを行ったことが，実際の相互評価をする際に役に立った

と答えた学生が多く，逆に役に立たなかったと答えた学生は一人もいなかったという質問３の回答結果からは，評価者トレーニングは，評価に役に立ったと考えられる。

表 56 に，質問 4. 評価者としての経験は，あなたにとってどのような影響を与えましたか？の回答結果を示す。

表 56

質問 4. 評価者としての経験は，あなたにとってどのような影響を与えましたか？

（件数）

	到達目標が明確になった	客観的に見られるようになった	その他
詳細グループ	4	5	3
簡易グループ	4	4	4
合計	8	9	7

自由記述の回答からは，大きく分けると，「到達目標が明確になった」と「客観的に見られるようになった」という２つのカテゴリーに分かれた。具体的には，「客観的に見られるようになった」というカテゴリーにおいては，「他の人を評価することで、自分が周りにどう映っているのかを知ることができ、発表者として求められることも分かった。」，「他人の意見、主張、発表の仕方の良い点、改善点を客観的に捉えることができるようになった。」，「人の発表を評価することで、自分の発表について客観的に見ることができるようになった。」などがあった。「到達目標が明確になった」というカテゴリーにおいては，「どういう発表をすれば良いのかがわかる」，「自分の発表の時に気をつけるポイントなどが分かるようになった。」，「評価することによって、より発表者に集中され自分のプレゼンで活かせるような要素を発見できたと思う」などの回答があった。

表 57 に，質問 5. 学生が評価することのメリット，デメリットは何ですか？の回答結果を示す。

表 57

質問 5. 学生が評価することのメリット，デメリットは何ですか？

（件数）

	メリット	デメリット
詳細グループ	10	7
簡易グループ	10	11
合計	20	18

学生が評価することのメリットとして，「改善点や発表の良し悪しを知ることで，自分の発表に役立つ」，「多数の評価があることで，公平性につながる」，「同年代の人に評価される点」，「評価基準により自分の発表が良くなる」などがあった。一方，デメリットとしては，「厳しい評価をしにくい」，「評価に偏りやばらつきが出る」，「評価経験が浅いため，自分の評価が正しいか不安」，「発表者との関係によって，評価が変更される」，「評価が甘くなる」，「私情を挟んでしまう」など評価の公平性や信頼性に関するものが多かった。また，運営面のデメリットとして，「プレゼンを聴いている間，評価ばかりに気を取られる」との意見もあった。メリットとしては，すでに学生による相互評価の利点として述べられているものとおおむね合致する。

9. 全体の考察

分析１では，評価者トレーニングが学生の評価力に影響を及ぼすかを調べるため，中間発表の前に評価者トレーニングを１回行い，中間発表に対する学生による評価と教員による評価の相関係数を求め分析を行った。その結果，評価者トレーニングを本格的に行った詳細グループにおいても，簡易的に行った簡易グループにおいても，評価の合計点において，弱い相関が得られたことから，両グループの評価力は低いことがわかった。つまり，評価者トレーニングが学生の評価力を上げるということを証明できなかった。

学生による相互評価と教員による評価の相関係数を評価項目別に見てみると，「内容」，「スライド」，「発音」の項目においては，簡易グループより詳細グループの相関が高くなったが，「オリジナリティ」と「アイコンタクト」の項目では，詳細グループより簡易グループの相関が高くなった。評価者トレーニングは，評価者の評価力を高めるために行われるものであることから，評価者トレーニングを行った後の学生の評価力は，詳細グループの評価力が，簡易グループの評価力より高くなるものと予想していた。しかし，評価項目別には差が出たが，合計では優劣は出なかった。つまり，一度の評価者トレーニングの質や内容の違いでは，学生の評価力に影響を及ぼすとは言えない。

１回目の評価者トレーニング後の詳細グループの評価力があまり伸びなかった背景にある理由として，これまで学生が相互評価の際に使用していた評価シートには，評価項目しか与えられておらず，判定基準についての詳細は記載されていないものを使用していたことがあげられる。そのため，評価項目と判定基準が明確に記されたルーブリックを用いて評価する経験は，学生にとっては初めての経験であったことから，トレーニングを受けたことで，逆に評価の際に混乱が生じた可能性がある（鈴木, 2005）。このように，学生の評価力には評価経験の浅さが大きく影響をしていると考えられる（笠巻, 2018; Weigle, 1994; 山西, 2004）。

評価経験の浅い学生に対して，評価者トレーニングを行ったことが，かえって学生に評価の際の混乱を招いたかどうかを調べるため，評価者トレーニングに対する学生の反応を調べるためにアンケート調査を行ったところ，筆者が作成したルーブリックはとてもわかりやすく，また評価者トレーニングは実際の評価の際に大変役立ったなど，好意的な反応が得られた。この結果から，筆者が懸念していた評価者トレーニングが相互評価の際に学生を混乱させたということはないと言えよう。

評価者トレーニングの１回目を行った後の中間発表の評価に対する学生の評価力が，詳細グループにおいても，簡易グループにおいても高くないことから，１回の評価者トレーニングでは，学生の評価力に影響を及ぼさないという分析１の結果から，分析２では，最終発表の前に評価者トレーニングの２回目を行い，最終発表に対する学生の評価力を調べ，評価者トレーニングの回数が学生に評価力に影響を及ぼすかを調べた。その結果，詳細グループも簡易グループも評価の合計点において，１回目の評価者トレーニング後の教員による評価と学生による相互評価の相関係数は弱い相関だったにもかかわらず，２回目の評価者トレーニング後では中程度の相関が得られたことから，両グループの全体的な評価力が１回目の評価者トレーニング後より高くなったことが確認された。つまり，評価者トレーニングの１回目から２回目にかけて，学生の評価力が伸びたと言える。この結果から，評価者トレーニングの回数を重ねることで，学生の評価力に影響を及ぼす可能性があると言えよう。

また，２回目の評価者トレーニング後の全体的な評価力においては，両グループの間に差は見られなかったが，評価項目別に見ると差が見られた。「アイコンタクト」，「発音」「ポーズ」の項目においては，簡易グループより詳細グループの評価力が高いことがわかった。つまり，これら３つの評価項目に対する評価力には，２回の評価者トレーニングが影響を及ぼした可能性があると考えられる。しかし，「内容」においては，詳細グループより簡易グループの評価力が高いことがわかり，また，「オリジナリティ」と「スライド」の項目では，両グループの間に差は見られなかった。評価者トレーニングを２回行えば，本格的な評価者トレーニングを行った詳細グループのすべての評価項目に対する評価力が，簡易グループより高くなると予想されたが，そのよ

うにはならなかった。

さらに，各グループにおける評価者トレーニングの１回目後の評価と，２回目後の評価の差を調べ，学生の評価力の伸びを調べた。その結果，詳細グループにおいては，評価の合計と「内容」を除くすべての評価項目において，伸びが確認された。一方，簡易グループにおいては，評価の合計，また「内容」，「スライド」において伸びが見られたが，「アイコンタクト」，「発音」においては，伸びが見られなかった。また，「オリジナリティ」と「ポーズ」においては，２回目の評価が１回目の評価よりも低くなっており，評価力が下がっていることがわかった。

分析１および２においても，評価項目の「内容」に対する評価力は，本格的な評価者トレーニングを２回行ってまでしても，高めることはできなかった。これは，「内容」の評価には，発表内容に対する評価者の予備知識や，発表内容を十分理解できるだけの英語力といった他の要因が関係していると考えられる。今後，それらの要因が及ぼす影響を検証する必要がある。

二回にわたる本格的な評価トレーニングを行った詳細グループと，簡易的にトレーニングを行った簡易グループとの間には，合計点においては、学生による評価と教員により評価の間の相関係数にはに差は見られなかったが，評価項目別に見た場合は両グループの評価力に差が見られた。このことは，一度の評価者トレーニングの質や内容の違いでは，学生の評価力に影響を及ぼすとは言えなかったが，二度のトレーニングでは，その質や内容の違いが，評価項目別に見た学生の評価力に影響を及ぼしたと考えられる。

笠巻 (2016)と笠巻 (2018)の研究結果から，学生にとって，発表に関する評価項目は評価しやすい一方，内容に関する項目は評価しにくいことがわかったが，二回の評価者トレーニングを行っても，内容に関する評価力を上げることができなかった。上で述べた原因の他に，評価者トレーニングの実施方法が挙げられる。評価者トレーニングは，ルーブリックと過去の学生の発表動画を用いて，明示的に指導していくことが有効である(Hughes, 1989)にもかかわらず，すべての評価項目において，そのようには指導できなった。特に「内容」と「オリジナリティ」においては，詳細グループの評価者トレーニングにおいても，筆者によるルーブリックを用いた口頭説明にとどまっていた。これらの項目を明示的に指導できる方法を考える必要がある。今後の課題としたい。

また，相互評価を行う際，クラスメイトの発表を聴くのと同時に評価を行うことは，評価経験の少ない学生にとってある程度の負担があることが想像される。しかも，そこにルーブリックを見ながら評価を行うという作業は，さらに負担が大きくなってしまった可能性がある。これについては，当日に発表を控えている学生は評価から外すなど，発表時の相互評価の実施方法も見直す必要があるであろう。

最後に，評価者トレーニングを行うことは，ただ単に教員による評価との一致の度合いをあげるためだけではないと筆者は考える。トレーニングを行うことにより，学生に評価力がついてくる，つまり，プレゼンテーションの良し悪しがわかるようになるのである。このように学生の批判的思考が高まることは，将来学生が自律した学習者になるために必要な判断力を身に着けることにつながる。また，評価者トレーニングを行い，学生にとって評価が難しい評価項目を知ることは，教員自身による評価をも見直す良い機会ともなり得る。つまりは，学生の評価力をあげるためにだけでなく，教員の評価力をあげることにもつながると言えよう。

注

1. この項目は，手違いで，詳細グループのみに尋ねた。

第 7 章　研究 5 :「評価項目別評価者トレーニング」が評価者としての学生の評価力に影響を及ぼすか？

1.　はじめに

第 6 章の結果から，評価者トレーニングは学生の評価力を上げるのに有効な方法であることがわかったが，実際の教育現場で，授業中に本格的な評価者トレーニングを実施することは, 時間の関係上, 難しいという欠点もある。第 6 章の研究においても，本格的な評価者トレーニングを行うのに要した時間は 45 分間であったため, 実施できたのは，実際に相互評価を行う授業日の前週に行うに留まった。しかし，中間発表と最終発表，それぞれの前週に評価者トレーニングを２回ずつ行った結果から，所要時間は短いが簡易な評価者トレーニングを２回行うよりも，時間はかかるが詳細な評価者トレーニングを行うことで，さらに学生の評価力を上げることができることがわかった。

評価者トレーニングは，評価項目と判定基準が明確に書かれたルーブリックと過去の学生の発表動画を用いて，明示的に指導していくことが有効である（Hughes,1989）ことから，研究 5 では，評価者トレーニングを指導の一環として，授業に組み入れて実施した。１回の評価者トレーニング所要時間を短縮し，到達目標である評価項目を指導する度に評価者トレーニングを行った。そして，実際の相互評価の前週まで，３回評価者トレーニングを実施した。

参加者は大学１年生で，プロジェクト発信型英語プログラムでの初めてのプレゼンテーションであること，相互評価の経験がない，あるいは浅いこと，また，これまでの研究で, 発表内容における学生の評価力が他の評価項目よりかなり低いことを鑑み, 評価項目については，筆者の判断で評価項目を「発表」のみに絞った。

研究 5 は，評価者トレーニングの実施方法の違いによって，プレゼンテーションに対する学生による相互評価と教員による評価との相関が変わるかを調べ，評価項目別評価者トレーニングが学生の評価力に影響を及ぼすかを検証する。

2.　研究 5 の目的

研究 5 では，指導の一環として指導する評価項目別評価者トレーニングが学生の評価力に影響を及ぼすか 調べることを目的とする。

3.　研究 5 における指導と評価方法

3.1　指導

指導内容と授業の流れについては，p. 10, p.11 で述べたので，ここでは省略する。

3.2　評価方法

評価方法については，p. 11, p.12 で述べたので，ここでは省略する。

ただし，研究 5 では，評価項目を「発表」に絞り，この項目を細分化し，「姿勢」，「アイコタンタクト」，「ジェスチャー」，「声の大きさ」，「明瞭さ」，「スライド」とした。

4.　研究方法

4.1　参加者

研究 5 の参加者は，滋賀県内の私立大学における筆者担当クラスの 1 年生 4 クラス計 79 名である。参加者の TOEIC IP テストの平均点は 464 点である。

4.2　分析対象者

参加者 79 名のうち，何らかの事情で，学生一人ひとりのパフォーマンスに対する学生による評価の点数を記入していない学生 3 名を除いた 76 名を分析対象とした。

評価項目別評価者トレーニングを行う 2 クラス（43 名）を詳細グループとし，相互評価を行う前週の授業時に 1 度のみ全評価項目に対する評価者トレーニングを行う 2 クラス（36 名）を簡易グループとした。

1 回目の評価者トレーニング後の中間発表における学生による評価と教員による評価の合計の相関係数が同じになるようにして，両グループの評価力を等質にし，詳細グループ（20 名），簡易グループ（20 名）の計 40 名を最終的な分析対象者とした。

表 1 と表 2 に 1 回目の評価者トレーニング後の中間発表における学生による評価と教員による評価の「合計」の相関係数の記述統計量と検定結果を示す。

表 1

「合計」の記述統計量

	n	*M*	*SD*	最大値	最小値
詳細グループ	20	.699	.094	.852	.478
簡易グループ	20	.698	.093	.850	.477

表 2

「合計」のマン・ホイットニーの *U* 検定結果・効果量

	U	*W*	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
詳細グループ							
VS	194.5	404.5	-.148	.888	-.02	なし	なし
簡易グループ							

表１と表２から，評価の合計のにおける教員による評価と学生による評価の相関係数は，詳細グループと簡易グループの間に差がないことがわかる。つまり，両グループの評価力は等質であると言える。

4.3　研究 5 における評価項目別トレーニング

4.3.1　評価トレーニングを行う前の準備

4.3.1.1　ルーブリック

研究５では，上述の評価項目に基づいて，ルーブリックを作成した。

図１に，研究５の評価者トレーニングで使用したルーブリックを示す。

P1 中間発表 Rubrics

注意事項 ・10段階評価 ・迷ったら，間の数字を使う ・小数点は使わない		The Physical Messages					The Visual Message
		Posture (姿勢)	アイコンタクト	ジェスチャー	Volume (声の大きさ)	Clarity (明瞭さ)	スライド
ほぼ完璧なレベル→	10	良い姿勢と適度な体の動きをしながら、発表している。	完全にオーディエンスを見て発表している	ジェスチャーを適切に使っている	何を言っているか、よくわかる大きさ	何を言っているか、はっきりとよくわかる	とても見やすく、伝えたい内容がよくわかる
	8	ほぼ良い姿勢と適度な体の動きをしながら、発表している。	ほぼオーディエンスを見て話している	ほぼ適切にジェスチャーを使っている	何を言っているのか、ほぼわかる大きさ	何を言っているのか、ほぼわかる	たまにわかりにくい箇所があるが、ほぼ伝えたい内容がわかる
	6	必要以上に体を動かしたり、下を向いたり、スクリーンをみたりしながら発表している。	だいたいオーディエンスを見て話している	ジェスチャーを使ったり、使わなかったりと半々	何を言っているのか、わかる時とわからない時と半々	何を言っているのか、わかる時とわからない時と半々	わかりやすいスライドとわかりにくいスライドが半々
	4	悪い姿勢（例：ほとんど下を向いていたり、スクリーンの方をみていたり）で発表していることの方が多い	あまりオーディエンスを見て話していない	あまりジェスチャーを使っていない	声が大きくなくて、何を言っているのか、あまりよくわからない	何を言っているのか、あまりよくわからない	何が伝えたいのか、あまりよくわからない
実質上の最低レベル→	2	悪い姿勢（例：ほとんど下を向いていたり、スクリーンの方をみていたり）で発表している	全くオーディエンスを見て話していない	ほとんどジェスチャーを使っていない	声が小さくて、何を言っているのか、わからない	何を言っているのか、わからない	何が伝えたいのか全くわからない
	0	発表者はプレゼンテーションをしていない					

図 1. 評価者トレーニングで使用したルーブリック

4.3.1.2　ルーブリックの判定基準を説明するためのスライドのサンプル

研究 5 では，スライドのサンプルは，第 6 章の研究で用いたものと同じものを使用したため，ここでは省略する。

4.3.1.3　評価者トレーニングで使用するサンプル・ビデオ

研究 5 で使用するサンプル・ビデオは，第 6 章の研究で用いたものと同じであるため，ここでは省略する。

4.3.2　実施手順

研究 5 では，中間発表の前に評価者トレーニングを行った。

次の表 3 のようなスケジュールで評価者トレーニングを行った。

表 3

評価者トレーニングの流れ

	指導する評価項目	評価者トレーニング	
Week1		詳細グループ	簡易グループ
Week2			
Week3			
Week4	姿勢，アイコンタクト，ジェスチャー	評価項目別評価者トレーニング	―
Week5	声の大きさ，明瞭さ	評価項目別評価者トレーニング	―
Week6	スライド	評価項目別評価者トレーニング	全評価項目まとめて評価者トレーニング
Week7	中間発表		
Week8	中間発表		

詳細グループに対し，授業で指導する評価項目ごとに， week4（「姿勢」，「アイコンタクト」，「ジェスチャー」），week5（「声の大きさ」，「明瞭さ」），そして week6（「スライド」）の中間発表の前の３週間をかけて，評価者トレーニングを行った。簡易グループに対しては，中間発表の１週間前の授業時にのみ，全評価項目に対する評価者トレーニングを１回行った。評価者トレーニングの手順を表 4 に示す。

表 4

評価者トレーニングの手順

	手順	詳細グループ	簡易グループ
1	学生にルーブリックと評価シートを配布	○	○
2	ルーブリックに書かれている評価項目および評価基準の中から，その週に指導および評価者トレーニングを行う項目についての説明を行う。	○	—
2	ルーブリックに書かれている全ての評価項目および判定基準について，一度に説明を行う。	—	○
3	「姿勢」，「アイコンタクト」，「ジェスチャー」，「声の大きさ」，そして「明瞭さ」において筆者が実演する	○	○
4	「スライド」にはサンプルを用いる。	○	○
5	サンプル動画の中から一人（A 君）のビデオを見せる	○	○
6	A 君の評価をさせる。	○	○
7	評価シートを回収する。	○	○
8	その週に評価者トレーニングを行う評価項目において，判定基準の各レベルごとに筆者が実演し，説明を行う。	○	—
8	全評価項目の判定基準の各レベルごとに一度に筆者が実演し，説明を行う。	—	○
9	サンプル動画の中から二人（B 君，C 君）のビデオを見せる	○	○
10	評価をさせる	○	○
11	教員から B 君，C 君の改善点についてコメントする	○	○
12	評価シートを回収する	○	○
13	新たに評価シートを配布する	○	○
14	再度，二人（B 君，C 君）のビデオを見せる	○	○
15	評価をさせる	○	○

17	評価シートを回収する	○	○
18	教師による，ルーブリックに対応した適切と思われる評価点を提示する。	○	○
19	新たに評価シートを配布する	○	○
20	再度，A 君のビデオを見せる	○	○
21	評価をさせる	○	○
22	評価シートを回収する	○	○
23	教師による，ルーブリックに対応した適切と思われる評価点を提示する。	○	○

詳細グループにおいて，1 回の評価項目別評価者トレーニングに要した時間は約 30 分で 3 回で合計 90 分であった。一方，簡易グループのトレーニングに要した時間は約 45 分であった。

5. 分析：評価項目別評価者トレーニングが学生の評価力に影響を及ぼすか

5.1 分析方法

教員は優れた発表には高い評価をつけ，あまり良くない発表には低い評価をしているが，学生が正しく評価できるかどうかを調べるために，学生もまた教員による評価と同じように，優れた発表には高い評価を，あまりよくない発表には低い評価をしているかを調べた。学生が行ったプレゼンテーションに対する学生による相互評価と，教員による相互評価との相関係数を詳細グループ，簡易グループごとに算出した。学生一人ひとりのパフォーマンスに対する学生による評価の合計点の正規性の検定を行った。その結果を表 5 に示す。正規性が認められなかったことと，担当クラスの数の関係でデータ数が少ないことにより，スピアマンの相関係数を用いた。次に，2 群の相関係数に差があるかを調べるために，マン・ホイットニーの *U* 検定を行い，効果量を算出した。

表 5

正規性の検定結果

	Kolmogorov-Smirnov(a)		
	統計量	自由度	有意確率
詳細グループ	.045	841	.000
簡易グループ	.053	611	.000

6. 結果

6.1 群別の分析結果と考察

表 6 に，評価項目別に評価者トレーニングを行った詳細グループにおける，学生による相互評価と教員による評価との相関係数の記述統計量を示す。図 2 は，詳細グループの箱ひげ図である。

表 6

詳細グループの記述統計量 （*n*=20）

	M	*SD*
姿勢	.278	.207
アイコンタクト	.672	.109
ジェスチャー	.614	.124
声の大きさ	.656	.124
明瞭さ	.430	.148
スライド	.197	.176
合計	.699	.094

表 6 が示すように，詳細グループにおいては，評価の合計点において中程度の相関が得られた（*r* = .699）。しかし，評価項目別に見ると，アイコンタクト（*r* = .672），ジェスチャー（*r* = .614），声の大きさ（*r* = .656），明瞭さ（*r* = .430）においては中程度の相関が得られたが，姿勢（*r* = .278）は弱い低い相関，スライド（*r* = .170）においては相関がないことがわかった。

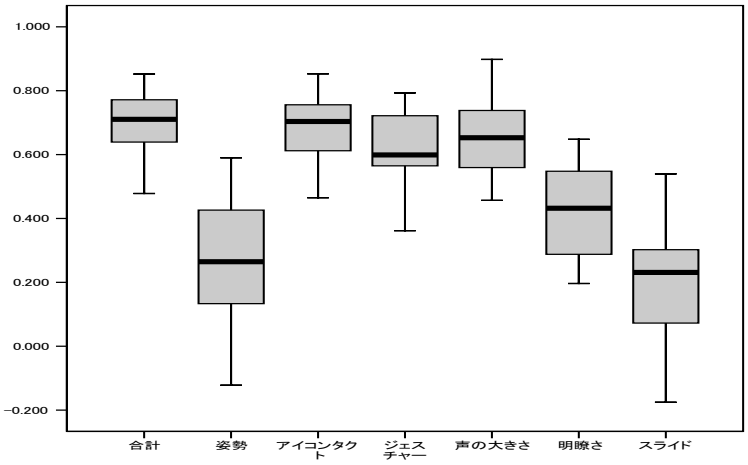


図 2. 詳細グループの箱ひげ図

図 2 からは，評価の合計（*SD* = .094）においては，高い相関（*r* = .852）から中程度の関係（*r* = .472）になるなど，相関係数が 2 段階にばらついていることがわかる。また，評価項目別に見ると，「アイコンタクト」（*SD* = .109）は高い相関（*r* = .853）から中程度の相関（*r* = .465），「声の大きさ」（*SD* = .124）は，高い相関（*r* = .898）から中程度の相関（*r* = .457）になるなど，それぞれの相関係数が 2 段階にばらついていることがわかった。つまり，これらの項目においては，詳細グループの評価力はある程度高いと言える。

また，「明瞭さ」（*SD* = .148）においては，中程度の相関（*r* = .648）から相関がない関係（*r* = .196），「ジェスチャー」（*SD* = .124）は，高い相関（*r* = .793）から低い相関（*r* = .362）になるなど，相関係数が 3 段階にばらついていることがわかった。つまり，これらの項目においては，詳細グループの評価力は高くないと言える。

さらに，「姿勢」（*SD* = .207）は中程度の相関（*r* = .590）からマイナスの関係（*r* = −.122），「スライド」（*SD* = .176）中程度の相関（*r* = .540）からマイナスの関係（*r* = −.175）になるなど，相関係数が 4 段階に大きくばらついていることがわかった。つまり，これらの項目においては，詳細グループの評価力は低いと言える。

これらの結果から，評価項目別に評価者トレーニングは，評価項目によっては学生の評価力に影響を与えないことがわかった。

表 7 に，相互評価を行う前週の授業においてのみ評価者トレーニングを行った簡易グループにおける，学生による相互評価と教員による評価との相関係数の記述統計量を示す。

表 7

簡易グループの記述統計量　（*n*=20）

	M	*SD*
姿勢	.458	.129
アイコンタクト	.623	.150
ジェスチャー	.665	.199
声の大きさ	.398	.195
明瞭さ	.234	.231
スライド	.307	.244
合計	.698	.093

簡易グループにおいては，評価の合計点において，中程度の相関が得られた（*r* = .698）。しかし，評価項目別に見ると，アイコンタクト（*r* = .623），ジェスチャー（*r* = .665），姿勢（*r* = .458）においては中程度の相関が得られたが，声の大きさ（*r* = .398），

明瞭さ（$r = .234$），スライド（$r = .307$）においては低い相関が得られた。

図 3 に，簡易グループの箱ひげ図を示す。

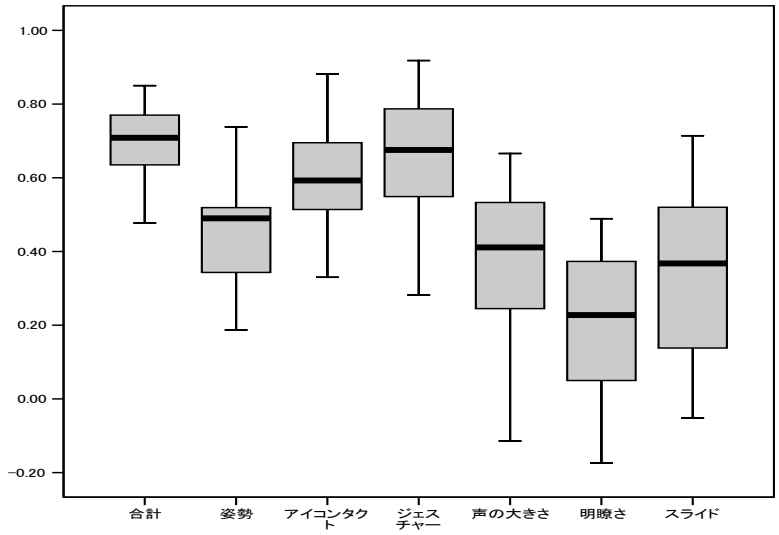


図 3. 簡易グループの箱ひげ図

図 3 から，評価の合計（$SD = .093$）においては，高い相関（$r = .850$）から中程度の関係（$r = .477$）になるなど，相関係数が 2 段階にばらついていることがわかる。しかし，評価項目別に見ると，「アイコンタクト」（$SD = .150$）は，高い相関（$r = .882$）から低い相関（$r = .330$），さらに「ジェスチャー」（$SD = .199$）においても，高い相関（$r = .918$）から低い相関（$r = .282$）になるなど，それぞれの相関係数が 3 段階にばらついていることがわかった。また，「姿勢」（$SD = .129$）は高い相関（$r = .738$）から相関がない関係（$r = .187$），「声の大きさ」（$SD = .195$）は，中程度の相関（$r = .666$）からマイナスの関係（$r = -.014$），「明瞭さ」（$SD = .231$）も，中程度の相関（$r = .631$）からマイナスの関係（$r = -.122$）になるなど相関係数が 4 段階に大きくばらついており，さらに「スライド」（$SD = .176$）においては，高い相関（$r = .764$）からマイナスの関係（$r = -.066$）になるなど，相関係数が 5 段階に大きくばらついていることがわかった。これらの結果から，簡易グループの評価力は高くないことがわかった。

6.2　評価項目別の群間の分析結果と考察

6.1 で，学生の総合的な評価力は，詳細グループ，簡易グループのどちらも高くないことが確認された。そこで，詳細グループ，簡易グループそれぞれの学生の評価力が，評価項目によって違いがあるかを調べた。次に，詳細グループと簡易グループの相関

係数に差があるかを調べるために，マン・ホイットニーの *U* 検定によって多重比較を行った。さらに効果量を算出した。

表 8 と表 9 に「姿勢」における教員による評価と学生による評価の相関係数の記述統計量と検定結果を示す。

表 8

評価項目「姿勢」の記述統計量

	n	*M*	*SD*	最大値	最小値
詳細グループ	20	.278	.207	.590	-.122
簡易グループ	20	.458	.129	.738	.187

表 9

評価項目「姿勢」のマン・ホイットニーの *U* 検定結果・効果量

	U	*W*	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
詳細グループ VS 簡易グループ	85	295	-3.111	.001	-.492	中	簡易グループ

表 8 と表 9 から，評価項目「姿勢」における教員による評価と学生による評価の相関係数は，詳細グループと簡易グループの間に差があり，効果量は中（r = -.49）で，「姿勢」の項目においては，簡易グループの方が評価力が高いことがわかった。このことから，「姿勢」に対する評価力に，評価項目別評価者トレーニングは影響を及ぼさなかった可能性がある。

このような結果となった原因として，筆者が作成したルーブリックに書かれた「姿勢」の評価基準の文言に，「良い姿勢」と「適切な体の動き」というのがあることが考えられる。これらの文言では，評価者の判断にその基準が左右されてしまうことになりかねない。そのため，姿勢に関する指導を行った直後に評価者トレーニングを行う評価項目別評価者トレーニングをしても，実際の評価の際には評価者自身の判断基準で評価を行ったしまった可能性がある。

続いて，表 10 と表 11 に「アイコンタクト」における教員による評価と学生による評価の相関係数の記述統計量と検定結果を示す。

表 10

評価項目「アイコンタクト」の記述統計量

	n	*M*	*SD*	最大値	最小値
詳細グループ	20	.672	.109	.853	.465
簡易グループ	20	.623	.150	.882	.330

表 11

評価項目「アイコンタクト」のマン・ホイットニーの *U* 検定結果・効果量

	U	*W*	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
詳細グループ VS 簡易グループ	155	365	-1.217	.231	-.193	小	詳細グループ

表 10 と表 11 から，評価項目「アイコンタクト」における教員による評価と学生による評価の相関係数は，詳細グループと簡易グループの間に有意差はないが，効果量が小（*r* = -.19）で差が見られ，「アイコンタクト」の項目においては，詳細グループの方が評価力が高いことがわかった。

また表 10 が示すように，教員による評価と学生による評価の相関係数は，詳細グループ，簡易グループともに中程度であることから，発表中にアイコンタクトがしっかりとれているかどうかの判断は学生にとってわかりやすく，評価しやすい項目であると考えられる。そこでさらに，アイコンタクトの取り方の指導を受けた直後に，評価者トレーニングを行ったことで，「アイコンタクト」に対する評価力に，評価項目別評価者トレーニングは少しではあるが影響を及ぼした可能性がある。

次に，表 12 と表 13 に「ジェスチャー」における教員による評価と学生による評価の相関係数の記述統計量と検定結果を示す。

表 12

評価項目「ジェスチャー」の記述統計量

	n	*M*	*SD*	最大値	最小値
詳細グループ	20	.614	.124	.793	.362
簡易グループ	20	.665	.199	.918	.282

表 13

評価項目「ジェスチャー」のマン・ホイットニーの *U* 検定結果・効果量

	U	*W*	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
詳細グループ VS 簡易グループ	140	350	-1.623	.108	-.257	小	簡易グループ

表 12 と表 13 から，評価項目「ジェスチャー」における教員による評価と学生による評価の相関係数は，詳細グループと簡易グループの間に有意差はないが，効果量が小（*r* = -.25）で差が見られ，「ジェスチャー」の項目においては，簡易グループの方が評価力が高いことがわかった。このことから，「ジェスチャー」に対する評価力に，評価項目別評価者トレーニングは影響を及ぼさなかった可能性がある。

「ジェスチャー」の項目に対するこれらの結果にも，「姿勢」と同じ原因があると考えられる。「姿勢」同様，ルーブリックの判定基準に「適切に」という文言を用いていることから，具体性に欠き，ジェスチャーの良し悪しの判断基準は評価者の判断基準によるものになった可能性がある。このため，ジェスチャーの指導をした直後に評価項目別評価者トレーニングを行っていても，実際の評価の際には，評価者の判断基準が大きく影響してしまったと考えられる。

次に，表 14 と表 15 に「声の大きさ」における教員による評価と学生による評価の相関係数の記述統計量と検定結果を示す。

表 14

評価項目「声の大きさ」の記述統計量

	n	*M*	*SD*	最大値	最小値
詳細グループ	20	.656	.124	.898	.457
簡易グループ	20	.398	.195	.666	-.114

表 15

評価項目「声の大きさ」のマン・ホイットニーの *U* 検定結果・効果量

	U	*W*	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
詳細グループ VS 簡易グループ	45	255	-4.193	.000	-.663	大	詳細グループ

表 14 と表 15 から，評価項目「声の大きさ」における教員による評価と学生による評価の相関係数は, 詳細グループと簡易グループの間に差があり, 効果量は大（*r* = -.66）で,「声の大きさ」の項目においては，詳細グループの方が評価力が高いことがわかった。発表者本人にとって，発表の際，どの程度の声の大きさで話したら良いかというのはわかりにくいものである。そこで，声の大きさについて，授業で指導を受けた後すぐに，サンプル・ビデオを見て，評価者トレーニングを行ったことで，具体的にどの程度の声の大きさが適切なのかということがわかったと考えられる。このことから,「声の大きさ」に対する評価力に，評価項目別評価者トレーニングは大きな影響を及ぼした可能性がある。

次に，表 16 と表 17 に「明瞭さ」における教員による評価と学生による評価の相関係数の記述統計量と検定結果を示す。

表 16

評価項目「明瞭さ」の記述統計量

	n	*M*	*SD*	最大値	最小値
詳細グループ	20	.430	.148	.648	.196
簡易グループ	20	.234	.231	.631	-.174

表 17

評価項目「明瞭さ」のマン・ホイットニーの *U* 検定結果・効果量

	U	*W*	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
詳細グループ VS 簡易グループ	95	305	-2.840	.004	-.45	中	詳細グループ

表 16 と表 17 から，評価項目「明瞭さ」における教員による評価と学生による評価の相関係数は，詳細グループと簡易グループの間に差があり，効果量は中（r = -.45）で，「明瞭さ」の項目においては，詳細グループの方が評価力が高いことがわかった。「明瞭さ」においても「声の大きさ」同様，学生にとって，自分の話し方が明瞭であるかどうかの判断は難しいと思われる。そのため，どういう話し方をすれば，自分が伝えたいことがオーディエンスによくわかるのかについて指導を受けた直後に，サンプルビデオを用いた評価者トレーニングをしたことで，より明確にわかったと考えられる。このことから，「明瞭さ」に対する評価力に，評価項目別評価者トレーニングは中程度の影響を及ぼした可能性がある。

次に，表 18 と表 19 に「スライド」における教員による評価と学生による評価の相関係数の記述統計量と検定結果を示す。

表 18

評価項目「スライド」の記述統計量

	n	*M*	*SD*	最大値	最小値
詳細グループ	20	.197	.176	.540	-.175
簡易グループ	20	.307	.244	.714	-.066

表 19

評価項目「スライド」のマン・ホイットニーの U 検定結果・効果量

	U	*W*	*Z*	*p*	効果量 *r*	効果の大きさ	優劣関係
詳細グループ VS 簡易グループ	147	357	-1.434	.157	-.227	小	簡易グループ

表 18 と表 19 から，評価項目「スライド」における教員による評価と学生による評価の相関係数は，詳細グループと簡易グループの間に有意差はないが，効果量は小（r = -.22）で差が見られ，「スライド」の項目においては，簡易グループの方が評価力が高いことがわかった。スライドは，その良し悪しが見た目にわかりやすく，学生にとって評価しやすいものであると考えられる。しかし，スライドには内容が反映されていることから，発表内容をよく理解できていないと，スライドの良し悪しを適切に評価しにくい項目とも言えよう。そして，ルーブリックにも書かれているように，「スライド」の判定基準は見た目の良し悪しだけではなく，発表者が伝えたい内容が伝わったかどうかも含めて判断しなければならないため，発表内容の理解の程度に影響され

た可能性がある。このことから,「スライド」の評価力には，評価項目別評価者トレーニングより，内容理解という他の要因が影響を及ぼしたと思われる。また，内容を理解するためには，十分な英語力が必要とされることから，詳細グループの中には，クラスメイトの発表内容を十分理解するだけの英語力がない学生や，スライドの見た目に引きずられて評価をした学生や，また，発表内容に対する予備知識があまりない学生が混在していた可能性も考えられる。

上述の結果を表 20 にまとめた。

表 20
分析結果のまとめ

	姿勢	アイコンタクト	ジェスチャー	声の大きさ	明瞭さ	スライド
詳細グループ VS 簡易グループ	簡易グループ	詳細グループ	簡易グループ	詳細グループ	詳細グループ	簡易グループ

指導する評価項目ごとに評価者トレーニングを行う評価項目別評価者トレーニングは，実際の相互評価の直前にのみ行う評価者トレーニングと比べると，学生の評価力をより高めるものと予想していたが，表 19 が示すように，そのような結果にはならなかった。つまり，評価者トレーニングの実施方法の違いだけでは，学生の評価力に影響を及ぼすとは言えない。影響を及ぼす可能性のある要因については，考察で述べる。

7. 全体の考察

6.1 では，評価項目別に評価者トレーニングを行うことで学生の評価力に影響を及ぼすかを調べた。詳細グループにおいては，中間発表の前の３週間の授業時において，指導する評価項目ごとに評価者トレーニングを行い，簡易グループにおいては，中間発表の前週の授業時にのみ詳細な評価者トレーニングを１回行った。中間発表に対する学生による相互評価と教員による評価を相関係数を求めて分析を行った結果，詳細グループ，簡易グループともに，評価の合計点においては，中程度の相関が得られて，学生の評価力に影響を与えることはできた。しかし，項目別にみると，すべての項目に効果があったわけではなかった。

学生による評価と教員による評価の相関係数を評価項目別に見てみると，詳細グループにおいては，「アイコンタクト」，「ジェスチャー」，「声の大きさ」と「明瞭さ」においては中程度の相関が得られたが，「姿勢」においては弱い相関が得られ，さらに「スライド」においては相関がない関係となった。簡易グループにおいては，「姿勢」，「ア

イコンタクト」,「ジェスチャー」においては中程度の相関が得られ,「声の大きさ」,「明瞭さ」そして「スライド」においてはそれぞれ弱い相関が得られた。研究 5 では,評価経験の浅い学生にとっても評価しやすいとされる発表に関する評価項目に絞って評価者トレーニングを行ったため,すべての評価項目において,学生による評価と教員による相関は高くなるものと思われたが,そのような結果にはならなかった。

評価項目別に評価者トレーニングを行った詳細グループの評価力が,項目によって,評価者トレーニングの効果が見られた項目とそうでない項目があった原因の一つに,筆者が作成したルーブリックに書かれた評価基準の文言が考えられる。学生による評価と教員による評価の相関が著しく低かった「姿勢」と「ジェスチャー」は,見た目にその良し悪しがわかりやすく,学生にとって評価しやすいものと思われたが,「姿勢」と「ジェスチャー」の判定基準の文言の中に,「適切な」や「適度な」といった具体性を欠いた曖昧な表現を用いてしまったことで,良し悪しの基準を,結果として評価者である学生に委ねてしまった可能性がある。これについては,筆者の反省すべき点であり,文言を見直す必要がある。今後の課題としたい。

また,「姿勢」,「アイコンタクト」,および「ジェスチャー」の３項目に対する評価者トレーニングを実施したのは中間発表当日から数えると３週間も前のことになるため,仮にトレーニング実施直後では判定基準が明確に理解できていたとしても,時間が経つにつれて,基準が明確でなくなってしまい,評価者の判断基準になってしまったのではないかと考えられる。「姿勢」と「ジェスチャー」の項目において,簡易グループの相関係数が高い理由も,簡易グループに対しては,一度ではあるが,中間発表の直前に評価者トレーニングを行ったため,判定基準がぶれずに評価できた可能性がある。

また,評価項目の「明瞭さ」と「声の大きさ」の違いが学生にはよくわかっていなかった可能性も考えられる。

6.2 では,学生による評価と教員による評価の相関係数を評価項目別に検討したが,「アイコンタクト」,「声の大きさ」と「明瞭さ」においては,簡易グループより詳細グループの相関が高く,「姿勢」,「ジェスチャー」と「スライド」の項目においては,詳細グループより簡易グループの相関が高いことがわかった。研究 5 では,中間発表の前の授業において,指導する評価項目ごとに評価者トレーニングを３週間行う「評価項目別評価者トレーニング」の方が,実際に相互評価を行う直前に評価者トレーニングを１回のみ行うよりも,学生の評価力はすべての項目において高くなるものと思われたが,そのような結果にはならなかった。

このような結果になった原因の一つとして,トレーニングの実施回数が挙げられる。評価項目別評価者トレーニングは,上述したように,授業で指導する評価項目に沿って行ったため,実際の相互評価を行うまでに同じ項目のトレーニングを繰り返し実施

することができなかった。第７章の研究から，学生の評価力を上げる評価者トレーニングを２度行うことで，学生の評価力をより伸ばすことができることがわかっていたが，評価者トレーニングを評価項目別に分割したことで，結果として，各項目ごとに１度のトレーニングしかできなかったことで，学生の評価力をあげることができなかったことが原因のひとつであろう。

そのほかの要因として，筆者が教員として参加していたプロジェクト発信型英語プログラムでは，学生が行うプレゼンテーションに対する評価項目や判定基準についての説明は実際の発表の前週に行われた。そのため，学生は実際の発表の１週間前になって，ようやく到達目標を知ることになる。しかし，研究５で行った指導では，実際の発表時に用いられる評価項目および判定基準について，詳細グループ，簡易グループともに，発表の３週間前から説明および指導を受けていたため，学生自身が自分のプレゼンテーションを準備にするにあたり，到達目標が明確になって準備に臨めたことが考えられる。研究５における評価者トレーニングの内容の違いによって，両群の評価力の間にあまり大きな差が出なかった理由の一つと思われる。

最後に，評価者トレーニングは，学生の評価力を高めるだけではなく，プレゼンテーションの良し悪しがわかる判断力を身につけさせ，学生のプレゼンテーションの質を上げることにつながると思われる。また，指導と評価を一致させる評価項目別評価者トレーニング行うことで，学生に到達目標を明確にするだけではなく，教員自身の指導および評価を見直す良い機会ともなり得ると考える。

第8章　研究6：評価者トレーニングは学生のプレゼンテーション力に影響を及ぼすか？

1. はじめに

評価者トレーニングは, 評価者としての評価力をつけるために行われるものであり, 第 6 章，第 7 章の研究からも，評価者トレーニングは学生の評価の経験の浅さからくる評価力の無さを補い，それにより学生の評価力をあげることにつながることが検証された。しかし，評価者トレーニングを学生に実施することで，学生の「プレゼンテーション力」を高めることができることを実証した研究は，筆者が調べた限りでは見当たらない。評価力が高くなるということは，プレゼンテーションの良し悪しを自分で判断できるようになるということから，評価者トレーニングは，評価力を高めることだけではなく，学生のプレゼンテーションの成績，つまり「プレゼンテーション力」をも高めることができるのではないだろうかと考え，研究 6 では，評価者トレーニングが学生のプレゼンテーション力に影響を及ぼすかについて検証する。

2. 研究 6 の目的

研究 6 では，大学のスピーキング・クラスにおいて，評価者トレーニングを行ったトレーニングの実施内容の違いは，学生のプレゼンテーション力に影響するかどうかを調べることを目的に分析 1 と分析 2 を行った。分析 1 については 3.で，分析 2 については 4.で述べる。

分析 1：評価者トレーニングを行うことで，学生のプレゼンテーション力に影響を及ぼすか。また，プレゼンテーション力のどの分野に影響を及ぼすのか。
分析 2：評価者トレーニングの内容の違いは，プレゼンテーション力に影響を及ぼすか。また，プレゼンテーション力のどの分野に影響を及ぼすのか。

3. 分析 1: 評価者トレーニングを行うことで, 学生のプレゼンテーション力に影響を及ぼすか。また，プレゼンテーション力のどの分野に影響を及ぼすのか。

3.1 目的

分析 1 では，評価者トレーニングを受けた場合と，受けなかった場合とで，学生のプレゼンテーション力にどのように影響するかを調べることを目的とする。

3.2　研究方法

3.2.1　参加者

研究 6 の参加者は, 滋賀県内の私立大学における筆者担当クラスの 2015 年度 1 年生 61 名と，2016 年度後期 1 年生の 59 の計 120 名から，後述する方法によって選んだ 88 名を分析対象とした。

3.2.2　手続き

評価者トレーニングを行っていない 2015 年度の学生を「評価者トレーニングなしグループ」，評価者トレーニングを行った 2016 年度の学生を「評価者トレーニングありグループ」とする。これらのグループは，授業内容，授業回数，評価方法のいずれも同じである。ただし，評価項目については， 2015 年度の学生と 2016 年度の学生とで違うため，研究 6 では，両グループの共通の評価項目に絞って分析した。

両グループを等質にするために，両グループの学生の前期における中間発表の成績を用いて等質の 2 つのグループを作り，評価者トレーニングなしグループ 44 名，評価者トレーニングありグループ 44 名の計 88 名を分析対象とした。この 88 名のプレゼンテーションの成績において，評価者トレーニングを受けたグループと，評価者トレーニング受けなかったグループの間に差が生じるかを調べることにした。

3.3　指導と評価方法

3.3.1　指導

指導内容と授業の流れ，および評価者トレーニングの実施内容については第 6 章で述べたので，ここでは省略する。

3.3.2　評価方法

評価方法については，第 6 章で述べたので，ここでは省略する。

評価者トレーニングを行っていない 2015 年度の学生と評価者トレーニングを行った 2016 年度の学生とでは評価項目が異なるため, 研究 6 では, 両グループの共通の評価項目である,「リサーチ」,「オリジナリティ」と「発表」に絞った。

3.4　分析方法

評価者トレーニングを行った場合と，行わなかった場合とでは，学生のプレゼンテーション力に影響を及ぼすかを調べるために，両グループの後期における中間発表と最終発表の成績を比較した。２つのグループのプレゼンテーションの成績の伸びに差があるかを調べるために，マン・ホイットニーの U 検定を行い，効果量を算出した。

3.5 結果と考察

3.5.1 評価者トレーニングなしグループと評価者トレーニングありグループの等質性

両グループの学生の前期における中間発表の成績を用いて，両グループの学生のプレゼンテーション力が（統計的に）等質になるようにした。表 1 に前期の中間発表の成績の記述統計量を示す。

表 1
前期の中間発表の成績の記述統計量

評価項目	評価者トレーニング	*n*	*M*	*SD*
準備	なし	44	7.773	1.568
	あり	44	7.659	1.462
リサーチ	なし	44	7.364	1.399
	あり	44	7.523	1.532
オリジナリティ	なし	44	7.636	1.586
	あり	44	7.591	1.515
発表	なし	44	7.568	1.485
	あり	44	7.659	1.478
合計	なし	44	30.432	5.555
	あり	44	30.341	5.758

次に，学生一人ひとりの中間発表の成績の正規性の検定を行ったところ，表 2 の結果になった。

表 2
中間発表の成績の正規性の検定結果

評価項目	Kolmogorov-Smirnov(a)		
	統計量	自由度	有意確率
準備	.202	88	.000
リサーチ	.236	88	.000
オリジナリティ	.239	88	.000
発表	.204	88	.000
合計	.167	88	.000

表 2 から正規性が認められなかったため，検定の方法として，マン・ホイットニーの *U* 検定を行い，効果量を算出した。表 3 にマン・ホイットニーの *U* 検定結果・効果量を示す。

表 3
前期の中間発表の成績の検定結果及び効果量

評価項目	評価者トレーニング	平均ランク	順位和	*U*	*W*	*Z*	*p*	効果量 *r*	効果の大きさ
準備	なし	45.409	1998	928	1918	-.343	.737	-.037	無
	あり	43.591	1918						
リサーチ	なし	43.273	1904	914	1904	-.469	.643	-.050	無
	あり	45.727	2012						
オリジナリティ	なし	44.659	1965	961	1951	-.061	.959	-.007	無
	あり	44.341	1951						
発表	なし	43.659	1921	931	1921	-.318	.753	-.034	無
	あり	45.341	1995						
合計	なし	44.193	1944.5	954.5	1944.5	-.114	.912	-.012	無
	あり	44.807	1971.5						

表 3 の効果量の有無から，3.2.2 で述べた手続きにより構成した両グループの前期の中間発表におけるプレゼンテーション力が等質であることが確認された。

3.5.2　評価者トレーニングなしグループと評価者トレーニングありグループの後期における中間発表の成績比較

両グループの後期の中間発表（評価者トレーニング 1 回目後）の成績を比較分析した。表 4 に後期の中間発表の成績の記述統計量を示す。

表 4

後期の中間発表（評価者トレーニング 1 回目後）の成績の記述統計量

評価項目	評価者トレーニング	*n*	*M*	*SD*
リサーチ	なし	44	7.864	1.488
	あり	44	7.409	1.369
オリジナリティ	なし	44	7.909	1.736
	あり	44	5.659	1.293
発表	なし	44	7.750	1.806
	あり	44	6.909	.963
合計	なし	44	23.523	4.708
	あり	44	19.977	2.582

次に，後期の中間発表の成績の正規性の検定を行ったところ，表 5 の結果になった。

表 5

後期の中間発表（評価者トレーニング 1 回目後）の成績の正規性の検定結果

	Kolmogorov-Smirnov(a)		
評価項目	統計量	自由度	有意確率
リサーチ	.213	88	.000
オリジナリティ	.194	88	.000
発表	.119	88	.004
合計	.101	88	.027

表 5 から，正規性が認められなかったため，検定の方法として，マン・ホイットニーの *U* 検定を用いて分析した。また，効果量を算出した。

表 6 にマン・ホイットニーの *U* 検定結果・効果量を示す。「効果の大きさ」の欄には，効果の有無と，効果が認められる場合には，どちらのグループの成績が上回っているかを示している。

表 6

後期の中間発表（評価者トレーニング１回目後）の検定結果及び効果量

評価項目	評価者トレーニング	平均ランク	順位和	*U*	*W*	*z*	*p*	効果量 *r*	効果の大きさ
リサーチ	なし	47.784	2102.5	823.5	1813.5	-1.247	.215	-.133	小
	あり	41.216	1813.5						なし
オリジナリティ	なし	58.977	2595	331	1321	-5.410	.000	-.577	大
	あり	30.023	1321						なし
発表	なし	51.102	2248.5	677.5	1667.5	-2.439	.014	-.260	小
	あり	37.898	1667.5						なし
合計	なし	53.886	2371	555	1545	-3.450	.000	-.368	中
	あり	35.114	1545						なし

表 4 と表 6 から，中間発表（評価者トレーニング１回目後）の成績の合計点は，評価者トレーニングを行ったグループと，評価者トレーニングを行っていないグループとの間に効果量中（r = -.36）で差があり，評価者トレーニングを行っていないグループの成績がトレーニングを行ったグループより上回っていることがわかる。

項目別に見ても，「リサーチ」においては，有意差はないが，効果量小（r = -.13）で差が見られ，「リサーチ」以外のすべての項目には差があり，「オリジナリティ」は効果量大（r = -.57）で，「発表」においては効果量小（r = -.26）で差が見られた。これらの結果から，評価者トレーニングを行っていないグループの成績が，トレーニングを行ったグループの成績より上回っていることがわかった。

このことから，評価者トレーニングを１度行っただけでは，プレゼンテーションの成績には影響を及ぼさない可能性が考えられる。つまり，評価者としての評価力を高めるための評価者トレーニングが，学生自身の発表にも影響を及ぼすためには，たった１度のトレーニングではその効果は現れない可能性がある。

3.5.3　評価者トレーニングなしグループと評価者トレーニングありグループの後期における最終発表の成績比較

次に，両グループの後期の最終発表（評価者トレーニング２回目後）の成績を比較分析した。表 7 に中間発表の成績の記述統計量を示す。

表 7

後期の最終発表（評価者トレーニング 2 回目後）の記述統計量

評価項目	評価者トレーニング	*n*	*M*	*SD*
リサーチ	なし	44	7.818	1.467
	あり	44	7.795	1.440
オリジナリティ	なし	44	8.182	1.451
	あり	44	5.841	1.738
発表	なし	44	7.977	1.455
	あり	44	6.778	1.294
合計	なし	44	23.977	4.146
	あり	44	20.415	3.490

次に，後期の最終発表の成績の正規性の検定を行ったところ，表 8 の結果になった。

表 8

後期の最終発表（評価者トレーニング 2 回目後）の成績の正規性の検定結果

	Kolmogorov-Smirnov(a)		
評価項目	統計量	自由度	有意確率
リサーチ	.148	88	.000
オリジナリティ	.134	88	.000
発表	.134	88	.001
合計	.098	88	.036

表 8 の結果から，正規性が認められなかったため，検定の方法として，マン・ホイットニーの *U* 検定を用いて分析した。また，効果量を算出した。

表 9 にマン・ホイットニーの *U* 検定結果・効果量を示す。

表9

後期の最終発表（評価者トレーニング2回目後）の検定結果及び効果量

評価項目	トレーニング	平均ランク	順位和	*U*	*W*	*z*	*p*	効果量 *r*	効果の大きさ
リサーチ	あり	44.148	1942.5	952.5	1942.5	-.132	.897	-.015	無
	なし	44.852	1973.5						
オリジナリティ	あり	59.034	2597.5	328.5	1318.5	-5.423	.000	-.579	大
	なし	29.966	1318.5						なし
発表	あり	54.045	2378	548	1538	-3.531	.000	-.377	中
	なし	34.955	1538						なし
合計	あり	54.364	2392	534	1524	-3.628	.000	-.387	中
	なし	34.636	1524						なし

表7と表9から，最終発表（評価者トレーニング２回目後）の成績の合計点は，評価者トレーニングを行ったグループと，評価者トレーニングを行っていないグループとの間に差があり，効果量中（r = -.38）で，評価者トレーニングを行っていないグループの成績が，トレーニングを行ったグループより上回っていることがわかる。

項目別に見ると，「オリジナリティ」に差があり，効果量大（r = -.57）で，評価者トレーニングを行っていないグループの成績が，トレーニングを行ったグループの成績より上回っていることがわかった。また，「発表」にも差があり，効果量中（r =- .37）で，評価者トレーニングを行っていないグループの成績が，トレーニングを行ったグループの成績より上回っていることがわかった。「リサーチ」の項目に差はなかった。

このことから，評価者トレーニングを２回行っても，学生のプレゼンテーションの成績に影響を及ぼす可能性は低いと言えよう。このような結果となった原因として，評価者トレーニングを行ったことにより，どのような発表が良い発表で，またどのような発表が良くない発表なのかについて,その判定基準と共に学んだことで，学生の意識は，主にクラスメイトの発表を「評価する」ことに向けられてしまい，自身の発表に活かすまでには至らなかったのではないかと考えられる。評価者トレーニングを学生のプレゼンテーションの成績に反映させていくためには，１度や２度ではなく，継続して評価者トレーニングを行い，プレゼンテーションの前には必ず評価者トレーニングを行う必要があると考えられる。

3.5.4　評価者トレーニングなしグループと評価者トレーニングありグループの後期における成績の伸びの比較

次に，評価者トレーニングを受けたグループと，評価者トレーニングを受けなかったグループとでは，プレゼンテーションの成績の伸びに差が生じるかを調べた。

表 10 に，両グループの中間発表と最終発表の成績の差の記述統計量を示す。

表 10

中間発表と最終発表の成績の差の記述統計量

評価項目	評価者トレーニング	*M*	*SD*
リサーチ	あり	.386	1.498
	なし	-.045	1.430
オリジナリティ	あり	.182	1.646
	なし	.273	1.318
発表	あり	-.131	1.111
	なし	.227	1.655
合計	あり	.438	2.966
	なし	.455	3.794

次に，両グループの中間発表と最終発表の成績の差の平均値について正規性の検定を行ったところ，表 11 の結果になった。

表 11

中間発表と最終発表の成績の差の正規性の検定結果

	Kolmogorov-Smirnov(a)		
評価項目	統計量	自由度	有意確率
リサーチ	.202	88	.000
オリジナリティ	.236	88	.000
発表	.239	88	.000
合計	.204	88	.000

表 11 から，正規性が認められなかったため，検定の方法として，マン・ホイットニーの *U* 検定を用いて分析した。また，効果量を算出した。

表 12 に，マン・ホイットニーの *U* 検定結果・効果量を示す。

表 12

中間発表と最終発表の成績の差の検定結果及び効果量 (*n*=44)

評価項目	トレーニング	*M*	*SD*	平均ランク	順位和	*U*	*W*	*Z*	*p*	効果量 *r*	効果の大きさ
リサーチ	あり	.386	1.498	48.307	2125.5	800.5	1790.5	-1.434	.153	-.153	小
	なし	-.045	1.430	40.693	1790.5						あり
オリジナリティ	あり	.182	1.646	43.477	1913	923	1913	-.387	.703	-.042	無
	なし	.273	1.318	45.523	2003						
発表	あり	-.131	1.111	41.864	1842	852	1842	-.973	.333	-.104	小
	なし	.227	1.655	47.136	2074						なし
合計	あり	.438	2.966	45.318	1994	932	1922	-.301	.766	-.033	無
	なし	.455	3.794	43.682	1922						

表 10 と 12 から，プレゼンテーションの成績の合計点では，評価者トレーニングありグループと評価者トレーニングなしグループの間における，中間発表から最終発表への成績の伸びに，差はないことがわかる。

しかし，評価項目別に見ると，「オリジナリティ」の項目には差がないが，「リサーチ」と「発表」の項目には，有意差はないが，効果量がそれぞれ小で差が見られ，「リサーチ」では，評価者トレーニングを行った学生の方が，評価者トレーニングを行っていない学生より成績が伸びていることがわかる。また，「発表」では評価者トレーニングを行っていない学生の方が，評価者トレーニングを行った学生より成績が伸びていることがわかる。これは，「リサーチ」には，評価者トレーニングの影響が少し及ぼした可能性があるが，「発表」には影響を及ぼさなかったと考えられる。

「プロジェクト発信型プログラム」では，学生が自身の興味・関心をもとに選んだテーマに基づいてリサーチをした成果を発表することになっているため，自分のリサーチと同じ内容の発表は他に一つとしてない。つまり，同じような内容の発表がないということは，他に見習える例がないということである。その結果，学生は自分自身のリサーチの成果をどのように発表すればいいか，模索しながら発表に臨むことになる。それが，評価者トレーニングを行うことで，評価項目の「リサーチ」に対する判定基準を知ったうえで発表に臨めたことが，成績の上昇に貢献した可能性がある。

3.5.5 評価者トレーニングなしグループの中間発表と最終発表の比較

次に，評価者トレーニングを行わなかった学生 44 名を対象に，中間発表の成績から最終発表の成績へ変化があったかを調べた。表 13 に，評価者トレーニングなしグルー

プの中間発表と最終発表の記述統計量を示す。

表 13

評価者トレーニングなしグループの中間発表と最終発表の記述統計量

評価項目	発表	*M*	*SD*
リサーチ	中間	7.864	1.488
	最終	7.818	1.467
オリジナリティ	中間	7.909	1.736
	最終	8.182	1.451
発表	中間	7.750	1.806
	最終	7.977	1.455
合計	中間	23.523	4.708
	最終	23.977	4.146

次に，正規性の検定を行ったところ，表 14 の結果になった。

表 14

評価者トレーニングなしグループの正規性の検定結果

	Kolmogorov-Smirnov(a)		
評価項目	統計量	自由度	有意確率
リサーチ	.202	88	.000
オリジナリティ	.236	88	.000
発表	.239	88	.000
合計	.204	88	.000

表 14 から，正規性が認められなかったため，検定の方法として，ウィルコクスンの符号付き順位和検定を用いて比較した。

表 15 にウィルコクスンの符号付順位和検定の記述統計を示す。

表 15

ウィルコクスンの符号付順位和検定記述統計

		n	平均ﾗﾝｸ	順位和
リサーチ	負の順位	13	14.231	185
	正の順位	13	12.769	166
	同順位	18		
	合計	44		
オリジナリティ	負の順位	11	11.636	128
	正の順位	15	14.867	223
	同順位	18		
	合計	44		
発表	負の順位	16	16.656	266.5
	正の順位	19	19.132	363.5
	同順位	9		
	合計	44		
合計	負の順位	19	19.184	364.5
	正の順位	21	21.690	455.5
	同順位	4		
	合計	44		

次に, 表 16 に評価者トレーニングなしグループの中間発表と最終発表の成績を比較した記述統計量・ウィルコクスンの符号付き順位検定結果・効果量を示す。

表 16

「評価者トレーニングなしグループ」の記述統計量・検定結果・効果量　(*n*=44)

評価項目	発表	*M*	*SD*	*z*	*p*	効果量 *R*	効果の大きさ
リサーチ	中間	7.864	1.488	-.248	.821	-.027	無
	最終	7.818	1.467				
オリジナリティ	中間	7.909	1.736	-1.244	.225	-.133	小
	最終	8.182	1.451				最終
発表	中間	7.750	1.806	-.820	.419	-.088	無
	最終	7.977	1.455				
合計	中間	23.523	4.708	-.615	.545	-.066	無
	最終	23.977	4.146				

表 15 と 16 から，評価者トレーニングなしグループの学生のプレゼンテーションの成績は，ウィルコクスンの符号付き順位検定結果とその有意確率，及び効果量から，合計点において，中間発表と最終発表の間に差がないことがわかる。順位を上げた参加者数は 21 名、順位を下げた参加者数は 19 名で、両者の差はほとんどなく，効果量もなかった。

項目別に見ると，「オリジナリティ」に有意差はないが，効果量が小（r = -.13）で差が見られ，最終発表の成績が中間発表の成績を上回っていることがわかる。順位を上げた参加者数は 15 名、順位を下げた参加者数は 11 名で、両者の差が少しみられ,効果量が小であることがわかる。

「リサーチ」では，順位を上げた参加者数は 21 名、順位を下げた参加者数は 19 名で、両者の差はほとんどなく，効果量もなかった。

「発表」でも，順位を上げた参加者数は 19 名、順位を下げた参加者数は 16 名で、両者の差はほとんどなく，効果量もなかった。「発表」に影響を及ぼさなかった理由として，この項目には，評価者トレーニングだけではなく，英語の発話トレーニングも併せて行う必要があると考えられる。どのような発表の仕方が良いのか，または悪いのかということを他の学生の発表を見て学ぶだけでなく，実際に自分の発表に向けて練習していくことで，上手く発表できるようになるのであろう。

これらの結果から，評価者トレーニングを行わなかった学生の成績は，中間発表と最終発表ではあまり変わらないと言えよう。つまり，評価者トレーニングを行わなかった場合の学生のプレゼンテーション力に変化はないと考えられる。

3.5.6　評価者トレーニングありグループの中間発表と最終発表の成績比較

次に，評価者トレーニングを行った学生 44 名を対象に，中間発表の成績から最終発表の成績へ変化があったかを調べた。表 17 に，評価者トレーニングありグループの中間発表と最終発表の記述統計量を示す。

表 17

評価者トレーニングありグループの中間発表と最終発表の記述統計量

評価項目	発表	*M*	*SD*
リサーチ	中間	7.409	1.369
	最終	7.795	1.440
オリジナリティ	中間	5.659	1.293
	最終	5.841	1.738
発表	中間	6.909	.963
	最終	6.778	1.294
合計	中間	19.977	2.582
	最終	20.415	3.490

次に，正規性の検定を行ったところ，表 18 の結果になった。

表 18

評価者トレーニングありグループの正規性の検定結果

	Kolmogorov-Smirnov(a)		
評価項目	統計量	自由度	有意確率
リサーチ	.236	88	.000
オリジナリティ	.147	88	.000
発表	.067	88	.020
合計	.057	88	.020

表 18 から正規性が認められなかったため, 検定の方法としてウィルコクスンの符号付き順位和検定を用いて比較した。

表 19 にウィルコクスンの符号付順位和検定の記述統計を示す。

表 19

ウィルコクスンの符号付順位和検定の記述統計

		n	平均ﾗﾝｸ	順位和
リサーチ	負の順位	13	18.615	242
	正の順位	23	18.435	424
	同順位	8		
	合計	44		
オリジナリティ	負の順位	16	13.875	222
	正の順位	16	19.125	306
	同順位	12		
	合計	44		
発表	負の順位	19	23.289	442.5
	正の順位	21	17.976	377.5
	同順位	4		
	合計	44		
合計	負の順位	17	20.176	343
	正の順位	24	21.583	518
	同順位	3		
	合計	44		

次に, 表 20 に評価者トレーニングありグループの中間発表と最終発表の成績を比較した記述統計量・ウィルコクスンの符号付順位検定結果・効果量を示す。

表 20

「評価者トレーニングありグループ」の記述統計量・検定結果・効果量　(*n*=44)

評価項目	発表	*M*	*SD*	*z*	*p*	効果量 *r*	効果の大きさ
リサーチ	中間	7.409	1.369	-1.472	.145	-.157	小
	最終	7.795	1.440				最終
オリジナリティ	中間	5.659	1.293	-.804	.429	-.086	無
	最終	5.841	1.738				
発表	中間	6.909	.963	-.439	.667	-.047	無
	最終	6.778	1.294				
合計	中間	19.977	2.582	-1.135	.261	-.121	小
	最終	20.415	3.490				最終

表 19 と表 20 から，評価者トレーニングありグループの学生のプレゼンテーションの成績は，合計点において有意差はないが，効果量が小で差が見られ，最終発表の成績が中間発表の成績を上回っていることがわかる。順位を上げた参加者数は 24 名、順位を下げた参加者数は 17 名で、効果量は小（*r* = -.12）であった。

項目別に見ると，「リサーチ」に有意差はないが，効果量が小で差が見られ，最終発表の成績が中間発表の成績をわずかに上回っていることがわかる。順位を上げた参加者数は 23 名、順位を下げた参加者数は 13 名で，効果量は小（*r* = -.15）であった。

「オリジナリティ」では，順位を上げた参加者数は 16 名、順位を下げた参加者数は 16 名で、効果量はなかった。

「発表」でも，順位を上げた参加者数は 21 名、順位を下げた参加者数は 19 名で、効果量はなしであった。

以上のことから，評価者トレーニングを行うことによって，プレゼンテーションの成績が大きく伸びるわけではなく，学生のプレゼンテーション力に小さな影響しか及ぼさないと考えられる。

このような結果になった原因として，次のことが考えられる。評価者トレーニングを行うことによって，評価項目と判定基準がよくわかり，到達目標がより明確になったことで，プレゼンテーションの成績が上がったのであろう。評価者トレーニングを行わなかったグループは，中間発表の経験や教員，クラスメイトからのフィードバックをもとに，学期中２回目となる最終発表に臨んだと思われるが，評価者トレーニングを行ったグループは，それら以外に，評価者トレーニングによって，自身のプレゼンテーションの改善点がより明確になり，２回目の発表の最終発表へ臨めた可能性がある。

また,「オリジナリティ」に差がなかったが，評価者トレーニングを行わなかったグループでは差が出ていることからも（表 16），この項目には，評価者トレーニングは影響を及ぼさないと言えよう。つまり,「オリジナリティ」には，学生一人ひとりの発表内容に対する教員の予備知識が関係していると考えられる。つまり，学生のテーマに左右されるということである。評価者トレーニングを行わなくてもこの項目の成績が伸びたのは，評価者トレーニングなしのグループの学生の発表内容は，教員の予備知識に影響されないものだったと考えられる。

さらに,「発表」にも差がなかったが，この項目に評価者トレーニングが影響を及ぼさなかった理由として，3.5.5 でも述べたとおり，この項目には，評価者トレーニングだけではなく，音読練習やシャドーイングなどの英語の音声トレーニングも併せて行う必要があると考えられる。発表の良し悪しを見て学ぶのではなく，発表の練習をすることで，自身の発表が上達していくものと考えられる。

4. 分析 2: 評価者トレーニングの内容の違いが, 学生のプレゼンテーション力に影響を及ぼすか。また，プレゼンテーション力のどの分野に影響を及ぼすか。

4.1 目的

分析 2 では，本格的な評価者トレーニングを受けた場合と，簡易的な評価者トレーニングを受けた場合とで，学生のプレゼンテーション力にどのように影響するかを調べることを目的とする。

4.2 研究方法

4.2.1 参加者

3.2.1 で述べた 2015 年度 1 年生 61 名から，後述の方法によって選んだ 44 名を分析対象とした。,

4.2.2 手続き

本格的な評価者トレーニングを行ったグループを「詳細グループ」とし，簡易なトレーニングを行ったグループを「簡易グループ」とする。これらのグループは，授業内容，授業回数，評価方法のいずれも同じである。両グループを等質にするために，両グループの前期における中間発表の成績の合計点を用いて等質にし，詳細グループ 22 名，簡易グループ 22 名の計 44 名を分析対象とした。そして，詳細グループと簡易グループの間に，最終発表の成績に差が生じるかを調べることにした。

4.3　指導と評価方法

4.3.1　指導

指導内容と授業の流れ，及びトレーニングの実施内容については第 6 章で述べたので，ここでは省略する。

4.3.2　評価方法

評価方法については，第 6 章で述べたので，ここでは省略する。

4.4　分析方法

評価者トレーニングの内容の違いが，学生のプレゼンテーション力に影響を及ぼすかを調べるために, 両グループの後期における中間発表と最終発表の成績を比較した。２つのグループのプレゼンテーションの成績の伸びに差があるかを調べるために，マン・ホイットニーの *U* 検定を行い，効果量を算出した。

4.5　結果と考察

4.5.1　詳細グループと簡易グループの等質性

両グループの学生の前期における中間発表の成績を用いて，両グループの学生のプレゼンテーション力が等質になるようにした。表 21 に中間発表の成績の記述統計量を示す。

表 21
前期の中間発表の成績の記述統計量

評価項目	グループ	*n*	*M*	*SD*
準備	詳細	22	7.682	1.393
	簡易	22	7.636	1.560
内容	詳細	22	7.727	1.486
	簡易	22	7.318	1.585
オリジナリティ	詳細	22	7.636	1.560
	簡易	22	7.545	1.503
発表	詳細	22	7.500	1.504
	簡易	22	7.818	1.468
合計	詳細	22	30.545	5.663
	簡易	22	30.318	5.575

学生一人ひとりの中間発表の成績の正規性の検定を行ったところ, 表 22 の結果にな

った。

表 22

中間発表の成績の正規性の検定結果

	Shapiro-Wilk		
評価項目	統計量	自由度	有意確率
準備	.887	44	.000
リサーチ	.893	44	.001
オリジナリティ	.839	44	.000
発表	.898	44	.001
合計	.906	44	.002

表 22 から，正規性が認められなかったことと，データ数が少ないことから，２つのグループのプレゼンテーションの成績の伸びに差があるかを調べるために，マン・ホイットニーの *U* 検定を行い，効果量を算出した。

表 23 に・マン・ホイットニーの *U* 検定結果・効果量を示す。

表 23

前期の中間発表の成績の検定結果・効果量

評価項目	グループ	平均ランク	順位和	*U*	*W*	*Z*	*p*	効果量 *r*	効果の大きさ
準備	詳細	22.705	499.5	237.5	490.5	-.109	.922	-.017	無
	簡易	22.295	490.5						
内容	詳細	24.364	536	201	454	-.992	.332	-.150	小
	簡易	20.636	454						詳細
オリジナリティ	詳細	22.818	502	235	488	-.170	.874	-.026	無
	簡易	22.182	488						
発表	詳細	21.227	467	214	467	-.674	.523	-.102	小
	簡易	23.773	523						簡易
合計	詳細	22.86	503.00	234	487	-.190	.856	-.029	無
	簡易	22.14	487.00						

表 23 の効果量の有無から，4.2.2 で述べた手続きにより構成した両グループの評価者トレーニング１回目後の中間発表の成績の合計点におけるプレゼンテーション力が

等質であることが確認された。

しかし「内容」と「発表」の項目においては有意差はないが，それぞれ効果量が小で差が見られ，「内容」では詳細グループが，「発表」では簡易グループの成績が上回っていること確認された。

4.5.2 詳細グループと簡易グループの評価者トレーニング１回目後の成績比較

次に，両グループの評価者トレーニング１回目後の成績を比較分析した。表 24 に評価者トレーニング１回目後の成績の記述統計量を示す。

表 24

評価者トレーニング１回目後の記述統計量

評価項目	グループ	*n*	*M*	*SD*
内容	詳細	22	7.364	1.364
	簡易	22	7.455	1.405
オリジナリティ	詳細	22	5.773	1.152
	簡易	22	5.545	1.438
アイコンタクト	詳細	22	5.818	2.085
	簡易	22	6.455	2.017
スライド	詳細	22	8.318	.945
	簡易	22	8.227	1.343
発音	詳細	22	6.409	1.182
	簡易	22	6.409	1.141
ポーズ位置	詳細	22	6.727	1.032
	簡易	22	6.909	1.109
合計	詳細	22	40.409	4.584
	簡易	22	41.000	5.563

次に，両グループの学生一人ひとりの評価者トレーニング１回目後の成績の正規性の検定を行ったところ，表 25 の結果になった。

表 25

評価者トレーニング１回目後の成績の正規性の検定結果

	Shapiro-Wilk		
評価項目	統計量	自由度	有意確率
内容	.882	44	.000
オリジナリティ	.898	44	.001
アイコンタクト	.948	44	.046
スライド	.895	44	.001
発音	.899	44	.001
ポーズの位置	.870	44	.000
合計	.954	44	.007

表 25 から，正規性が認められなかったことと，データが少ないことから，検定の方法として，マン・ホイットニーの *U* 検定を用いて分析した。また，効果量を算出した。

表 26 にマン・ホイットニーの *U* 検定結果・効果量を示す。

表 26

評価者トレーニング 1 回目後の検定結果及び効果量 (*n*=22)

評価項目	グループ	平均ランク	順位和	*U*	*W*	*Z*	*p*	効果量 *r*	効果の大きさ
内容	詳細	22.114	486.5	233.5	486.5	-.211	.833	-.032	無
	簡易	22.886	503.5						
オリジナリティ	詳細	23.659	520.5	216.5	469.5	-.614	.539	-.093	無
	簡易	21.341	469.5						
アイコンタクト	詳細	20.5	451	198	451	-1.047	.295	-.158	小
	簡易	24.5	539						簡易
スライド	詳細	22.614	497.5	239.5	492.5	-.061	.951	-.010	無
	簡易	22.386	492.5						
発音	詳細	22.795	501.5	235.5	488.5	-.161	.872	-.025	無
	簡易	22.205	488.5						
ポーズ位置	詳細	21.864	481	228	481	-.348	.728	-.053	無
	簡易	23.136	509						
合計	詳細	22.25	489.5	236.5	489.5	-.130	.897	-.020	無
	簡易	22.75	500.5						

表 24 と表 26 から，評価者トレーニング 1 回目を行った後の詳細グループと簡易グループの成績の合計点において差はないことがわかった。つまり，両グループのプレゼンテーション力に差はない。

項目別に見ると，「アイコンタクト」に有意差はないが，効果量小（*r* = -.15）で差が見られ，簡易グループの成績が詳細グループの成績を上回っていることがわかった。その他の項目において両グループの間に差はなかった。

このことから，評価者トレーニングを 1 回行っただけでは，トレーニングの内容の違いが，学生のプレゼンテーション力に影響を及ぼす可能性は低いと考えられることから，継続して評価者トレーニングを行う必要がこの結果からも確認された。

4.5.3 詳細グループと簡易グループの評価者トレーニング 2 回目後の成績比較

続いて，両グループの評価者トレーニング 2 回目後の成績を比較分析した。

表 27 に評価者トレーニング 2 回目後の成績の記述統計量を示す。

表 27

評価者トレーニング 2 回目後の記述統計量

評価項目	グループ	*n*	*M*	*SD*
内容	詳細	22	7.545	1.262
	簡易	22	8.045	1.588
オリジナリティ	詳細	22	6.136	1.833
	簡易	22	5.545	1.625
アイコンタクト	詳細	22	5.682	2.033
	簡易	22	6.636	2.279
スライド	詳細	22	8.045	1.558
	簡易	22	8.364	1.255
発音	詳細	22	5.773	1.445
	簡易	22	6.273	1.609
ポーズ位置	詳細	22	6.409	1.141
	簡易	22	7.045	1.527
合計	詳細	22	39.591	7.436
	簡易	22	41.909	6.711

次に，両グループの学生一人ひとりの評価者トレーニング 2 回目後の成績の正規性の検定を行ったところ，表 28 の結果になった。

表 28

評価者トレーニング 2 回目後の成績の正規性の検定結果

	Shapiro-Wilk		
評価項目	統計量	自由度	有意確率
内容	.932	44	.013
オリジナリティ	.890	44	.001
アイコンタクト	.936	44	.017
スライド	.878	44	.000
発音	.909	44	.002
ポーズの位置	.921	44	.005
合計	.978	44	.006

表 28 から，正規性が認められなかったことと，データが少ないことから，検定の方

法として，マン・ホイットニーの *U* 検定を用いて分析した。また，効果量を算出した。

表 29 にマン・ホイットニーの *U* 検定結果・効果量を示す。

表 29

評価者トレーニング 2 回目後の検定結果及び効果量 (*n*=22)

評価項目	グループ	平均ﾗﾝｸ	順位和	*U*	*W*	*Z*	*p*	効果量 *r*	効果の大きさ
内容	詳細	19.614	431.5	178.5	431.5	-1.526	.127	-.231	小
	簡易	25.386	558.5						簡易
オリジナリティ	詳細	24.750	544.5	192.5	445.5	-1.196	.232	-.181	小
	簡易	20.250	445.5						詳細
アイコンタクト	詳細	19.727	434	181	434	-1.454	.146	-.220	小
	簡易	25.273	556						簡易
スライド	詳細	21.136	465	212	465	-.726	.468	-.110	小
	簡易	23.864	525						簡易
発音	詳細	20.455	450	197	450	-1.084	.278	-.164	小
	簡易	24.545	540						簡易
ポーズ位置	詳細	19.705	433.5	180.5	433.5	-1.503	.133	-.227	小
	簡易	25.295	556.5						簡易
合計	詳細	20.636	454	201	454	-.964	.335	-.146	小
	簡易	24.364	536						簡易

表 27 と表 29 から，評価者トレーニングの 2 回目を行った後の，詳細グループと簡易グループのプレゼンテーションの成績の合計点において有意差はないが，効果量小で差が見られ，簡易グループの成績が詳細グループの成績を上回っていることがわかった。

項目別に見ると，すべての項目において有意差はないが，「オリジナリティ」は，効果量小で差が見られ，詳細グループが簡易グループの成績を上回っているが，その他の項目においては，いずれも効果量小で差が見られ，簡易グループの成績が詳細グループの成績を上回っていることがわかった。これらの結果から，本格的な評価者トレーニングより，簡易的な評価者トレーニングを行った方が，学生のプレゼンテーション力にプラスの影響を及ぼす可能性があると考えられる。

詳細グループのプレゼンテーション力が簡易グループの成績より低くなった原因の一つとして，評価者トレーニングを詳細に行ったことで，一度に多くのことを詳しく

学ぶことになり，その分，学んだことを理解し吸収したうえで，自身のプレゼンテーションに反映させることが，却って難しくなったのではないかと考えられる。

また,「オリジナリティ」の項目以外全ての評価項目において簡易グループの成績が詳細グループの成績を上回っているのに対し,「オリジナリティ」のみ，詳細グループの成績が簡易グループの成績を上回った原因として,「オリジナリティ」には評価者トレーニング以外の要素，つまり，上述のとおり，学生一人ひとりの発表内容に対する教員の予備知識が関係していると考えられる。詳細グループの学生の発表内容は，教員の予備知識に影響されないものであったことが考えられる。

4.5.4 詳細グループと簡易グループのプレゼンテーションの成績の伸びの比較

詳細グループと簡易グループとでは，プレゼンテーションの成績の伸びに差が生じるかを調べた。表 30 に両グループの中間発表と最終発表の成績の差の記述統計量を示す。

表 30
中間発表と最終発表の成績の差の記述統計量

評価項目	トレーニング	*M*	*SD*
内容	詳細	.182	1.435
	簡易	.591	1.563
オリジナリティ	詳細	.364	1.649
	簡易	.000	1.662
アイコンタクト	詳細	-.136	2.660
	簡易	.182	2.218
スライド	詳細	-.273	1.549
	簡易	.136	1.457
発音	詳細	-.636	1.093
	簡易	-.136	1.320
ポーズの位置	詳細	-.318	1.129
	簡易	.136	1.246
合計	詳細	-.818	6.269
	簡易	.909	5.554

次に，学生一人一人の中間発表と最終発表の成績の点数差の正規性の検定を行ったところ，表 31 の結果になった。

表 31

中間発表と最終発表の成績の差の正規性の検定結果

	Shapiro-Wilk		
評価項目	統計量	自由度	有意確率
内容	.882	44	.000
オリジナリティ	.898	44	.001
アイコンタクト	.948	44	.046
スライド	.895	44	.001
発音	.899	44	.001
ポーズの位置	.870	44	.000
合計	.954	44	.047

表 31 から，正規性が認められなかったことと，データが少ないことから，検定の方法として，マン・ホイットニーの *U* 検定を用いて分析した。また，効果量を算出した。

表 32 に，両グループの中間発表と最終発表の成績の差の記述統計量・マン・ホイットニーの *U* 検定結果・効果量を示す。

表 32

中間発表と最終発表の成績の差の検定結果・効果量 (*n*=22)

評価項目	トレーニング	M	SD	平均ランク	順位和	U	W	Z	p	効果量 r	効果の大きさ
内容	詳細	.182	1.435	20.886	459.5	206.5	459.5	-.854	.401	-.129	小
	簡易	.591	1.563	24.114	530.5						簡易
オリジナリティ	詳細	.364	1.649	23.409	515	222	475	-.480	.640	-.073	無
	簡易	.000	1.662	21.591	475						
アイコンタクト	詳細	-.136	2.660	22.205	488.5	235.5	488.5	-.155	.882	-.024	無
	簡易	.182	2.218	22.795	501.5						
スライド	詳細	-.273	1.549	21.341	469.5	216.5	469.5	-.610	.546	-.093	無
	簡易	.136	1.457	23.659	520.5						
発音	詳細	-.636	1.093	20.614	453.5	200.5	453.5	-1.019	.320	-.154	小
	簡易	-.136	1.320	24.386	536.5						簡易
ポーズの位置	詳細	-.318	1.129	20.068	441.5	188.5	441.5	-1.291	.206	-.195	小
	簡易	.136	1.246	24.932	548.5						簡易
合計	詳細	-.818	6.269	21.114	464.5	211.5	464.5	-.718	.480	-.109	小
	簡易	.909	5.554	23.886	525.5						簡易

表 32 から，プレゼンテーションの成績の合計点では，詳細グループと簡易グループの間に有意差はないが，効果量小で差が見られ，簡易グループが詳細グループを上回っていることがわかる。

評価項目別に見ると，「オリジナリティ」，「アイコンタクト」，「スライド」には差がないことがわかる。しかし，「内容」，「発音」，「ポーズの位置」には有意差はないが，効果量小で差が見られ，いずれも簡易グループが詳細グループを上回っていることがわかる。

このような結果になった原因として，上述したとおり，本格的な評価者トレーニングを行った詳細グループの学生よりも，簡易的な評価者トレーニングを行った簡易グループの学生の方が，評価項目や判定基準に集中しやすく，その分よく理解もできて，

自身の 2 回目のプレゼンテーションに反映できたのではないかと考えられる。

4.5.5　詳細グループの評価者トレーニング 1 回目後と 2 回目後の成績比較

本格的な評価者トレーニングを行った詳細グループの学生 22 名を対象に, 中間発表の成績から最終発表の成績へ変化があったかを調べた。表 33 に詳細グループの記述統計量を示す。

表 33
詳細グループの記述統計量 (*n*=22)

評価項目	プレゼンテーション	*M*	*SD*
内容	中間	7.364	1.364
	最終	7.545	1.262
オリジナリティ	中間	5.773	1.152
	最終	6.136	1.833
アイコンタクト	中間	5.818	2.085
	最終	5.682	2.033
スライド	中間	8.318	.945
	最終	8.045	1.558
発音	中間	6.409	1.182
	最終	5.773	1.445
ポーズの位置	中間	6.727	1.032
	最終	6.409	1.141
合計	中間	40.409	4.584
	最終	39.591	7.436

次に，正規性の検定を行ったところ，表 34 の結果になった。

表 34

詳細グループの正規性の検定結果

	Shapiro-Wilk		
評価項目	統計量	自由度	有意確率
内容	.912	44	.003
オリジナリティ	.912	44	.003
アイコンタクト	.948	44	.046
スライド	.904	44	.001
発音	.929	44	.009
ポーズの位置	.911	44	.002
合計	.978	44	.048

表 34 から，正規性が認められなかったので，検定の方法として，ウィルコクスンの符号付順位和検定を用いて比較した。

表 35 に，ウィルコクスンの符号付順位和検定の記述統計を示す。

表 35

ウィルコクスンの符号付順位和検定記述統計

		n	平均ﾗﾝｸ	順位和
内容	負の順位	8	9.187	73.5
	正の順位	10	9.75	97.5
	同順位	4		
	合計	22		
オリジナリティ	負の順位	8	5.75	46
	正の順位	8	11.25	90
	同順位	6		
	合計	22		
アイコンタクト	負の順位	7	13.214	92.5
	正の順位	12	8.125	97.5
	同順位	3		
	合計	22		
スライド	負の順位	8	10.375	83
	正の順位	8	6.625	53
	同順位	6		
	合計	22		
発音	負の順位	12	7.375	88.5
	正の順位	2	8.25	16.5
	同順位	8		
	合計	22		
ポーズの位置	負の順位	11	9.318	102.5
	正の順位	6	8.417	50.5
	同順位	5		
	合計	22		
合計	負の順位	10	9.2	92
	正の順位	8	9.875	79
	同順位	4		
	合計	22		

表 36 に，詳細グループの中間発表と最終発表の成績を比較した記述統計量・ウィルコクスンの符号付順位和検定結果・効果量を示す。

表 36

詳細グループの記述統計量・検定結果・効果量 (*n*=22)

評価項目	発表	*M*	*SD*	*z*	*p*	効果量 *r*	効果の大きさ
内容	中間	7.364	1.364	-.539	.639	-.082	無
	最終	7.545	1.262				
オリジナリティ	中間	5.773	1.152	-1.160	.277	-.175	小
	最終	6.136	1.833				最終
アイコンタクト	中間	5.818	2.085	-.102	.932	-.016	無
	最終	5.682	2.033				
スライド	中間	8.318	.945	-.794	.478	-.120	小
	最終	8.045	1.558				中間
発音	中間	6.409	1.182	-2.336	.021	-.353	中
	最終	5.773	1.445				中間
ポーズの位置	中間	6.727	1.032	-1.301	.242	-.197	小
	最終	6.409	1.141				中間
合計	中間	40.409	4.584	-.284	.798	-.043	無
	最終	39.591	7.436				

表 35 と表 36 から，本格的な評価者トレーニングを行った詳細グループのプレゼンテーションの成績は,順位を上げた参加者数が8名,順位を下げた参加者数が10名で,両者の差はほとんどないことからも効果量がないことことがわかる。つまり，詳細グループにおける，評価者トレーニング 1 回目後と 2 回目後のプレゼンテーションの成績は，合計点において差がないことがわかる。

項目別に見ると，「内容」では，順位を上げた参加者数が 10 名，順位を下げた参加者数が 8 名で，両者の差はほとんどなく，効果量がないことがわかる。つまり，この評価項目の評価者トレーニング 1 回目後と 2 回目後のプレゼンテーションの成績は差がないことがわかる。

しかし，「オリジナリティ」では，順位を上げた参加者数が 8 名，順位を下げた参加者数が 8 名で，両者の人数には差はないが，順位和と平均ランクに差があり，効果量は小（r = -.17）であった。以上のことから，「オリジナリティ」では，最終発表の成績が中間発表の成績を上回っていることがわかる。

「アイコンタクト」では，順位を上げた参加者数が 12 名，順位を下げた参加者数が 7 名で，両者の人数に少し差があるが，順位和と平均ランクに差はなく，効果量もな

しであった。つまり，この評価項目の評価者トレーニング 1 回目後と 2 回目後のプレゼンテーションの成績は差がないことがわかる。

一方,「スライド」でも，順位を上げた参加者数が 8 名，順位を下げた参加者数が 8 名で，両者の人数には差はないが，順位和と平均ランクに差があり，効果量は小（r = -.12）であった。以上のことから,「スライド」では，中間発表の成績が最終発表の成績を上回っていることがわかる。

さらに,「発音」では, 順位を上げた参加者が 2 名, 順位を下げた参加者が 12 名で, 効果量は中であった。このことから，中間発表の成績が最終発表の成績を，効果量中（r = -.35）で上回っていることがわかる。

また,「ポーズの位置」では，順位を上げた参加者が 6 名，順位を下げた参加者が 11 名で，効果量は小（r = -.19）であった。以上のことから,「ポーズの位置」でも, 中間発表の成績が最終発表の成績を上回っていることがわかる。

これらの結果から，本格的な評価者トレーニングは，学生のプレゼンテーション力に影響を及ぼすとは言えない。このような結果となった原因として，評価者トレーニングの効果自体が現れるのに，ある程度の時間を要する可能性が考えられる。また，トレーニングを行ったのは，実際の発表の直前週であったことから，自身の発表の直前に, 評価項目や判定基準の説明を受けても, それらを理解してすぐに実践するのは, まだプレゼンテーションの経験が浅い学生には難しかったと思われる。特に,「ポーズ」や「発音」は，学習者が注意をしただけで良くなるものではなく，適切なモデルを用いて指導をしたうえで，音読練習やシャドーイングなどのトレーニングが必要であると考える。

また，中間発表の成績が最終発表の成績を上回った項目においては，評価者トレーニング以外のことが影響することが考えられる。いずれの項目も，発表の仕方に関するものであることから，評価者トレーニングだけでなく，発話のトレーニングも併せて行う必要があると考えられる。

4.5.6　簡易グループの評価者トレーニング 1 回目後と 2 回目後の成績比較

簡易的な評価者トレーニングを行った簡易グループの学生 22 名を対象に, 中間発表の成績から最終発表の成績へ変化があったかを調べた。表 37 に簡易グループの記述統計量を示す。

表 37

簡易グループの記述統計量 (*n*=22)

評価項目	発表	*M*	*SD*
内容	中間	7.455	1.405
	最終	8.045	1.588
オリジナリティ	中間	5.545	1.438
	最終	5.545	1.625
アイコンタクト	中間	6.455	2.017
	最終	6.636	2.279
スライド	中間	8.227	1.343
	最終	8.364	1.255
発音	中間	6.409	1.141
	最終	6.273	1.609
ポーズの位置	中間	6.909	1.109
	最終	7.045	1.527
合計	中間	41.000	5.563
	最終	41.909	6.711

次に，正規性の検定を行ったところ，表 38 の結果になった。

表 38

簡易グループの正規性の検定結果

	Shapiro-Wilk		
評価項目	統計量	自由度	有意確率
内容	.912	44	.003
オリジナリティ	.906	44	.002
アイコンタクト	.959	44	.012
スライド	.899	44	.001
発音	.915	44	.003
ポーズの位置	.890	44	.001
合計	.972	44	.025

表 38 から，正規性がなかったので，検定の方法として，ウィルコクスンの符号付順位和検定を用いて比較した。

表 39 にウィルコクスンの符号付順位和検定の記述統計を示す。

表 39

ウィルコクスンの符号付順位和検定記述統計

		n	平均ﾗﾝｸ	順位和
内容	負の順位	5	10.3	51.5
	正の順位	13	9.192	119.5
	同順位	4		
	合計	22		
オリジナリティ	負の順位	8	8.5	68
	正の順位	8	8.5	68
	同順位	6		
	合計	22		
アイコンタクト	負の順位	7	10.643	74.5
	正の順位	11	8.773	96.5
	同順位	4		
	合計	22		
スライド	負の順位	8	7.312	58.5
	正の順位	8	9.688	77.5
	同順位	6		
	合計	22		
発音	負の順位	11	7.955	87.5
	正の順位	6	10.917	65.5
	同順位	5		
	合計	22		
ポーズの位置	負の順位	6	8.75	52.5
	正の順位	9	7.5	67.5
	同順位	7		
	合計	22		
合計	負の順位	7	12.5	87.5
	正の順位	13	9.423	122.5
	同順位	2		
	合計	22		

表 40 に，簡易グループの中間発表と最終発表の成績を比較した記述統計量・ウィルコクスンの符号付順位検定結果・効果量を示す。

表 40

簡易グループの記述統計量・検定結果・効果量 (*n*=22)

評価項目	発表	*M*	*SD*	*z*	*p*	効果量 *r*	効果の大きさ
内容	中間	7.455	1.405	-1.517	.132	-.229	小
	最終	8.045	1.588				最終
オリジナリティ	中間	5.545	1.438	.000	1.000	.000	無
	最終	5.545	1.625				
アイコンタクト	中間	6.455	2.017	-.485	.641	-.074	無
	最終	6.636	2.279				
スライド	中間	8.227	1.343	-.503	.674	-.076	無
	最終	8.364	1.255				
発音	中間	6.409	1.141	-.544	.602	-.083	無
	最終	6.273	1.609				
ポーズの位置	中間	6.909	1.109	-.440	.728	-.067	無
	最終	7.045	1.527				
合計	中間	41.000	5.563	-.655	.524	-.099	無
	最終	41.909	6.711				

表 39 と 40 から，簡易的な評価者トレーニングを行った簡易グループのプレゼンテーションの成績は，合計点において順位を上げた参加者が 13 名，順位を下げた参加者が 7 名で，両者の差が少し見られるが，順位和と平均ランクに差がなく，効果量はなかった。つまり，最終発表の成績と中間発表の成績に差はないと言える。

項目別に見ると，「内容」では，順位を上げた参加者が 13 名，順位を下げた参加者が 5 名で，効果量は小（r = -.22）であった。このことから，最終発表の成績が中間発表の成績を上回っていることがわかる。

「オリジナリティ」では，順位を上げた参加者が 8 名，順位を下げた参加者が 8 名で，効果量はなかった。つまり，「オリジナリティ」では，最終発表の成績と中間発表の成績に差はないと言える。

「アイコンタクト」では，順位を上げた参加者が 11 名，順位を下げた参加者が 7 名で，順位和と平均ランクに差はなく，効果量はなかった。つまり，「アイコンタクト」

でも，最終発表の成績と中間発表の成績に差はないと言える。

「スライド」では，順位を上げた参加者が 8 名，順位を下げた参加者が 8 名で，効果量はなかった。つまり，「スライド」でも，最終発表の成績と中間発表の成績に差はないと言える。

「発音」では，順位を上げた参加者が 6 名，順位を下げた参加者が 11 名で，順位和と平均ランクに差はなく，効果量はなかった。以上のことから，「発音」でも，最終発表の成績と中間発表の成績に差はないと言える。

「ポーズの位置」では，順位を上げた参加者が 9 名，順位を下げた参加者が 6 名で，効果量はなかった。つまり，「ポーズの位置」でも，最終発表の成績と中間発表の成績に差はないと言える。つまり，これらの項目においては，簡易グループにおける中間発表の成績と最終発表の成績に差はないと言える。

これらの結果から，本格的な評価者トレーニング同様，簡易的な評価者トレーニングもまた，学生のプレゼンテーション力に影響を及ぼすとは言えない。このような結果となった原因として，評価者トレーニングの効果自体が現れるのに，ある程度の時間を要する可能性が考えられる。詳細グループにおいても，簡易グループにおいても，トレーニングを行ったのは，実際の発表の直前週である。自身の発表の直前に，評価項目や判定基準の説明を受けても，それらを理解してすぐに実践するのは，まだプレゼンテーションの経験が浅い学生には難しかったと思われる。特に，「ポーズ」や「発音」は，学習者が注意をしただけで良くなるものではなく，適切なモデルを用いた，練習が必要であると考える。

5. 全体の考察

分析 1 では，評価者トレーニングが学生のプレゼンテーション力に影響を及ぼすかを調べるために，評価者トレーニングを行ったグループと，評価者トレーニングを行わなかったグループそれぞれの，後期における中間発表と最終発表の成績を比較し分析した。

3.5.4 では，両グループの間における中間発表から最終発表への成績の伸びに差がなかったことから，評価者トレーニングが学生のプレゼンテーション力に影響を及ぼす可能性は低いことが考えられた。

しかし，グループ別に分析した結果，3.5.5 から，評価者トレーニングを行わなかった場合，中間発表から最終発表への成績に変化は見られなかったが，3.5.6 からは，評価者トレーニングを行った場合，最終発表の成績が中間発表の成績を若干ではあるが上回っていることことがわかった。このことは，評価者トレーニングは学生のプレゼンテーション力に少し影響する可能性があることを示している。

評価者トレーニングを行ったグループの成績の伸びが，評価者トレーニングを行わ

なかったグループの成績の伸びと比べて大きく変わらなかった原因として次のことが考えられる。一つ目は，評価者トレーニングの実施時期である。研究 6 では，発表の直前週の授業時間内にトレーニングを行ったが，その効果が現れるのは，トレーニング直後ではなく，むしろしばらく時間を置いてからではないだろうかと考える。そのため，研究 6 を行った期間内には，その効果が確認できなかった可能性がある。長期間にわたって何度もプレゼンテーションを行う機会を与えて、学生たちが評価基準を意識しながら準備と練習を行えば、評価者トレーニングが学生のプレゼンテーション力を伸ばす可能性がある。この点が十分でなかった可能性があるため，今後，学生のプレゼンテーション力を継続して観察し，評価者トレーニングが学生のプレゼンテーション力に影響を及ぼす可能性を再検証したいと考えている。今後の課題としたい。

二つ目の理由として，トレーニングの方法と実施回数が考えられる。第 6 章の研究から，１回の評価者トレーニングでは学生の評価力に影響を及ぼさなかったが，本格的な評価者トレーニングの回数を重ねることで，学生の評価力に影響を及ぼす可能性が高いことが確認されたことから，詳細なトレーニングを継続的に行うことで，その効果を期待できると思われる。また，学生は，評価者トレーニングを相互評価のために行われているものと思うことから，自身の発表とは結びつけてトレーニングを受けていなかった可能性もあるであろう。この点においては，学生にアンケートを取り，評価者トレーニングを自分の発表と結び付けて受けていたかどうかについて確認する必要がある。評価者トレーニングを学生の発表に反映させていくためには，さらにトレーニングの方法とその実施回数と共に，自分自身の発表を評価させることが重要な要因になるであろう。

もう一つの理由として，評価項目が考えられる。「発表」の項目については，トレーニングの実施の有無だけではなく，併せて発話トレーニングも行うことで，学生のプレゼンテーション力に影響を及ぼす可能性が考えられる。具体的には，聞き手に読み聞かせることを意識した音読練習をさせることでプレゼンテーション力が高まっていくと考える。そして，実際にプレゼンテーションを行うことで，うまくできた点，また反対にうまくできなかった点を発表者自身も気づき振り返ることで，次回の発表に向けて改善して臨むことができると考えられる。また，発表後のクラスメイトや教員から得たフィードバックも，次の発表に活かされたことであろう。さらに，回数をこなしていくことで，これまでできなかったこともできるようになるなど，ある程度の「経験値」もまたプレゼンテーション力に影響を及ぼすのではないかと考える。

また，「オリジナリティ」については，3.5.6 で述べたとおり，評価者トレーニングではなく，学生一人一人の発表内容に対する，教員の予備知識が関係していると考えられる。つまり，学生の発表のテーマに大きく左右されてしまう可能性がある。しかし，「プロジェクト発信型プログラム」では，教員が指定したテーマではなく，学生一

人一人が自身の興味・関心のもとにリサーチした成果を発表することになっている。担当する英語教員がすべての学生の発表内容について何等かの知識を持つことは難しい。評価を行う教員の予備知識が学生の発表の評価に影響しないようにするためには，発表内容に対して専門知識を持った教員と併せて評価を行うか，あるいは，それを実施することが難しければ，「オリジナリティ」という項目は，評価項目の中からなくすことも検討する必要があるのではないかと考える。

分析 2 では，分析 1 で実証された評価者トレーニングが学生のプレゼンテーション力に影響を及ぼす可能性を受けて，さらに評価者トレーニングの実施内容の違いによって，学生のプレゼンテーション力に影響を及ぼすかを調べた。本格的な評価者トレーニングを行った詳細グループと，簡易的な評価者トレーニングを行った簡易グループそれぞれの，後期における中間発表と最終発表の成績を比較し分析した。4.5.4 では，成績の合計点においては２つのグループの間に差がり，簡易グループの方が詳細グループより，若干成績が伸びていることがわかった。

また，グループ別に分析した結果，4.5.5 から，詳細グループにおいては，中間発表の成績が最終発表の成績を上回っていることがわかり，また，4.5.6 からは，簡易グループにおいては，中間発表の成績と最終発表の成績の間に差は見られなかった。これらの結果から，本格的な評価者トレーニングよりも，むしろ簡易的な評価者トレーニングを行った方が，学生のプレゼンテーション力に影響を及ぼす可能性があると言える。

しかし，4.5.5 および 4.5.6 の結果から，分析 1 と同様に，評価者トレーニングが学生のプレゼンテーション力に影響を及ぼす可能性は，トレーニングの内容にかかわらず低いと考えられる。

このような結果となった原因として，次のことが考えられる。一つ目として，詳細な評価者トレーニングは，一度に多くのことを学び，また理解しなくてはならないため，学生にとっては，あまり詳細ではなく，簡潔にトレーニングを行う方が，トレーニングの内容が頭に残りやすかったか可能性が考えられる。この点においては，実施方法を見直したい。

もう一つの理由として，両グループの間に，「内容」，「発音」，および「ポーズの位置」といった発表の仕方に関する項目において差があり，いずれも，詳細グループより簡易グループの方が成績が伸びていることから，評価者トレーニングが発表に関する項目の成績に影響を与えるためには，トレーニングが詳細か，簡易かといった内容の違いだけではなく，上述のとおり，発表に関する評価項目には，評価者トレーニングと併せて発話のトレーニングを行う必要があるであろう。

最後に，評価者トレーニングを，学生としての評価者の評価力を高める効果と，学生自身の発表にも影響を与える指導としての効果を両立させる効果的な方法を，これ

からも模索していきたいと考えている。

第9章　終わりに

1.　本研究で明らかになったこと

本研究では，学生による相互評価の信頼性を検証し，スピーキングの内容を評価するにあたり，学生の視点を取り入れられるかどうかを検討した。

教員は優れた発表には高い評価をつけ，あまり良くない発表には低い評価をしているが，学生もまた教員による評価と同じように，優れた発表には高い評価を，あまりよくない発表には低い評価をしているかを調べた。そこで，学生が正しく評価できるかどうかを調べるために，教員による評価と学生による相互評価の相関係数を算出した。そして，相関が高い場合，低い場合それぞれにおいて，その背景にある原因，およびその理由を考察した。評価者としての学生の評価力には，学生の「プレゼンテーション力」，「英語力」，「予備知識の有無」，「評価者トレーニング」が影響を及ぼす要因であることが考えられた[1]ことから，本研究では6つのリサーチ・クエスチョンを立て，検証を行った。

RQ.1 学生の「プレゼンテーション力」が評価者としての学生の評価力に影響を及ぼすか。

RQ.2 学生の「英語力」が評価者としての学生の評価力に影響を及ぼすか。

RQ.3 学生の「予備知識」が評価者としての学生の評価力に影響を及ぼすか。

RQ.4 「評価者トレーニング」が評価者としての学生の評価力に影響を及ぼすか。

RQ.5 「評価項目別評価者トレーニング」が評価者としての学生の評価力に影響を及ぼすか。

RQ.6 評価者トレーニングは学生のプレゼンテーション力に影響を及ぼすか。

第3章では，学生の評価力に影響を与えるとされる4つの要因のうち，RQ.1の学生のプレゼンテーション力が学生の評価者としての評価力に影響を及ぼすかを調べた。その結果，プレゼンテーション力が一番と高いとされる上位群の学生の評価力が一番低いことが確認されたことから，学生のプレゼンテーション力は，評価者としての学生の評価力には影響を及ぼさない可能性が高いことがわかった。また，学生の評価傾向として，学生全体としては，教員の評価と比べると甘めの評価を行うことがわかったが，成績群別にみると，上位群の学生は低めの評価を行う傾向があり，逆に下位群の学生は高めの評価を行う傾向があることがわかった。その要因として，学生は評価を行う際に，自分自身のプレゼンテーションの出来と比較していることが考えられた。

第4章では，学生の評価力に影響を及ぼすと考えられる4つの要因の一つである，RQ.2の学生の英語力が評価者としての学生の評価力に影響を及ぼすかを調べた。その結果，英語力が高いとされる上位群ではなく，中位群の評価力が一番高いことが確認されたことから，学生の英語力は，評価者としての学生の評価力には影響を及ぼさな

い可能性が高いことがわかった。。また，学生の評価傾向として，学生全体として，教員による評価と比べると甘目の評価を行う傾向があることがわかった。その要因として，クラスメイトに対して評価することに気まずさを感じることから，高めの点数をつけてしまうといった心理的な要因や，他者との関係性で評価を変えてしまうといった評価者の性格が，評価に影響を及ぼしていると考えられた。

第 5 章では，学生の評価力に影響を及ぼすと考えられる 4 つの要因の一つである，RQ.3 の予備知識が評価者としての学生の評価力に影響を及ぼすかを調べた。その結果，学生の予備知識の多少は，学生の評価傾向に少しの影響を及ぼす可能性があることがわかった。また，学生の評価傾向として，教員による評価と比べると，学生による評価は甘目になることがわかった。その要因としては，クラスメイトを評価することに対する心理的な要因，学生の評価経験の少なさが関係していることが考えられた。

学生による相互評価を活用するためには，学生に評価力をつけ，評価の信頼性を高めることが必要であり，そのためには，評価項目，判定基準および評価の観点が明確に示されたルーブリックを用いて，評価者トレーニングを行うことが必要であると考えられた。

第 6 章では，学生の評価力に影響を及ぼすと考えられる 4 つの要因の一つである，RQ.4 の評価者トレーニングが評価者としての学生の評価力に影響を及ぼすかを調べた。その結果，たった一度の評価者トレーニングでは，学生の評価力に影響を及ぼすとは言えないが，トレーニングを二度以上行うことで，学生の評価力に伸びが見られたことから，評価者トレーニングは回数を重ねることで学生の評価力に影響を及ぼす可能性が高いことがわかった。つまり，評価者トレーニングを行うことにより，学生の評価力が向上し，学生による相互評価の信頼性が高くなる可能性があることから，さらにトレーニングを実施することで，将来的に学生の相互評価を成績の一部に組み入れられる可能性が示唆された。

また，評価者トレーニングを行うことにより，学生はどのようなプレゼンテーションを目指すべきなのかを明示的に指導できることから，評価者トレーニングは波及効果があると考えられた。

しかし，二回にわたる評価者トレーニングを行っても，「内容」と「オリジナリティ」という発表の内容に関する項目に対する学生の評価力を上げることができなかったという課題が見つかった。

第 7 章では，RQ.5 の指導する評価項目別に評価者トレーニングを行うことは，実際の評価の直前に 1 度行う評価者トレーニングよりも，評価者としての学生の評価力に影響を及ぼすかを調べた。その結果，評価の合計においては，学生による評価と教員による評価に中程度の相関が得られたが，評価項目別にみると，中程度の相関から相関のない関係までさまざまであったことがわかった。つまり，評価項目別評価者トレ

ーニングは，学生の評価力に影響を及ぼすとは言えないが，詳細グループ，簡易グループのそれぞれの評価の合計点において，中程度の相関が得られたことから，第 6 章の研究結果と同様に，評価者トレーニング自体は，学生の評価経験の少なさからくる評価力の無さを補い，それにより学生の評価力をあげることにつながることがわかった。

また，研究 5 で筆者が作成したルーブリックの「姿勢」，「ジェスチャー」，「声の大きさ」と「明瞭さ」の項目の判定基準に，曖昧であると思われる文言があったことで，評価項目別評価者トレーニングを行った学生の評価力を上げることができなかった可能性がある。

しかし，指導する評価項目ごとに評価者トレーニングを行うことは，学生の評価力をあげるためだけではなく，指導と評価の一致の観点から，到達目標をより明確にできたと思われる。このことから，評価項目別評価者トレーニングには波及効果があると考えられ，また指導の一つとして授業に組み入れられる可能性を示唆できた。

第 8 章では, RQ.6 の評価者トレーニングが評価者としての学生の評価力を高めるだけではなく，評価者トレーニングをプレゼンテーションの指導となり得るかを調べた結果，評価者トレーニングは学生のプレゼンテーション力にわずかに影響を及ぼす可能性があることがわかった。また，評価者トレーニングの内容の違いが，学生のプレゼンテーション力に影響を及ぼすかを調べた結果，簡易的な評価者トレーニングを行う方が学生のプレゼンテーション力に影響を及ぼす可能性があることが確認された。

これらの結果から，評価者トレーニングが学生のプレゼンテーション力に影響を及ぼすためには，評価項目および評価者トレーニングの方法，実施回数，および実施する時期を見直し，手順を簡易にしたトレーニングを継続的に実施し，さらに音声トレーニングも併せることで，学生のプレゼンテーションの指導の一つとして活用できる可能性が示唆された。

本研究では，学生による相互評価を成績の一部に組み入れることを目標として，学生の相互評価の信頼性を調べるべく，学生の評価力および評価力に影響を及ぼす原因を調べてきた。その結果, 筆者がこれまで担当してきた学生の評価力は一貫して低く，信頼性に欠けるものばかりであった。そして，評価力に影響を及ぼすと考えられた学生のプレゼンテーション力，英語力も影響を及ぼさないことがわかったことから，学生の評価力の低さは，相互評価の経験が少ない可能性が考えられた。そのため，学生は評価の基準を自分のプレゼンテーションの出来と比べて判断してしまう傾向があることが考えられた。このことから，自分より上手い発表には高めの評価を，逆に下手と感じた発表には低めの評価をしてしまうことから，教員による評価との相関が低くなったと考えられる。

そして，評価者トレーニングを行い，学生の評価力が上がるかどうかを調べたところ，学生の評価力の無さはトレーニングを行うことである程度補うことが可能であることがわかった。しかし，トレーニングを行っても，内容に関する評価項目に対する評価力を上げることはできなかった。これには，内容の理解には，英語力だけでなく，内容に対するスキーマをどの程度持っているかどうかが大きく関わっていると考えられた。

教員がテーマを設定して，それに基づいて学生がリサーチをした成果を発表するのであれば，ある程度そのテーマに対するスキーマは授業内外において蓄えることができたかもしれないが，「プロジェクト発信型英語プログラム」では，学生が自ら興味・関心のあることに基づいてリサーチを行った成果を発表するため，学生の発表内容は一人ひとり違い，また多岐にわたるため，発表内容に対するスキーマをすべてにおいて持っておくことは難しいと言える。この点を改善するため，評価の際に使用している評価シートの評価項目を教員用と学生用に分け，学生にはデリバリー中心，教員は内容や英語力の評価も含めるとするなどが考えられる。

最後に，評価者トレーニングを行うことは，ただ単に教員による評価の一致の度合いをあげるためだけではないと筆者は考える。トレーニングを行うことにより，学生に評価力がついてくる，つまり，プレゼンテーションの良し悪しがわかるようになるのである。そして，クラスメイトの良い発表からも，そうでない発表からも学ぶことができるようになる。このように学生の批判的思考が高まることは，将来学生が自律した学習者になるために必要な判断力を身に着けることにつながる。これは，ただオーディエンスとしてクラスメイトの発表を聴いているだけでは得ることができないと言えよう。

また，評価者トレーニングを行い，学生にとって評価が難しい評価項目や，学生の評価傾向を知ることは，教員自身による評価をも見直す良い機会ともなり得る。つまりは，学生の評価力をあげるためにだけでなく，教員の評価力をあげることにもつながると言えよう。相互評価をただ単に評価の一つとして捉えるのではなく，相互評価を行うことが学びにつながるよう，今後も引き続き研究をしていきたい。

2. 本研究から示唆できること

学生の発表に対して，信頼性の高い評価を行うためには，まず評価者の数を増やし，可能であれば，学生の発表内容に詳しい評価者を用意することが必要であろう。しかし，実際には，学生の発表に対する評価は担当教員一人で行うことが多い現状から，担当外の教員と一緒に評価をすることは，容易なことでないと考えられる。そのような場合は，発表の「内容」に対する判定基準を，内容の良し悪しに関するものではなく，わかりやすさ，つまりは初めて発表を聴く人にもわかるように発表できているか

どうかにするのが良いであろう。

「プロジェクト発信型英語プログラム」は，学生が主体的に取り組めるプログラムのため，学生のモチベーションがとにかく最後まで落ちることがないことから，とても素晴らしい教授法の一つであると実感している。しかし，そこで行われている評価方法については，改善の余地があると筆者は考えている。具体的には，学生側においては，学生による相互評価の評価項目の見直し，および評価力のトレーニングの実施であり，教員側においては，判定基準をプログラムの主旨に従って明確にしたルーブリックの作成，また担当教員間における評価者間信頼性の確保である。これらを行っていくことにより，このプログラムの到達目標が教員間においても，学生教員間においても明確になることで，どの教員が担当しても，同様のプレゼンテーションができ，さらにそこに学生のオリジナリティが加わった他に類を見ない発表ができる学生が増えていくことにつながっていくと思われる。

もう一つの改善点として，学生のプレゼンテーション力を高めるためには，他に英語力向上のための科目との連携が必要であろう。この点においては，筆者が参加していた英語プログラムでは，実際には英語力向上のための科目が週 1 回 90 分で併設されていたが，研究 6 の結果から見ても，残念ながらその連携はうまく行っていなかったのではないかと考えられる。「プロジェクト発信型プログラム」では，これまで中学・高校で培ってきた英語力で，自分の興味関心に基づいてリサーチした成果を発信するという方針だが，学生がいま持てる英語力をさらに向上させるべく，教材を用いて，それをあらゆる角度から，繰り返しインプット，インテイクし，アウトプットにつなげる指導が必要である。

3. 今後の課題

本研究の結果から，上述したことが明らかになったが，具体的にどの要因が評価者としての学生の評価力に影響を及ぼすかということを特定することは難しいという結論に至った。その要因について明確に突き止めるには至っていない。

また，評価力を高めるための評価者トレーニングを指導の一つとして確立させるためにはさらなる研究が必要であると考えている。以下に筆者が今後取り組むべき課題についてまとめる。

1. 評価者としての学生の評価力に影響を及ぼすと考えられる他の要因について，新たに他の要因を探るだけでなく，考えられる複数の要因を組み合わせて研究していく必要があると考える。

2. 本研究では，評価者トレーニングにおいて，「内容」と「オリジナリティ」という発表の内容に関する項目の説明に，ルーブリックを示すことと，教員による口頭説明のみしかできなかったために，二回にわたる評価者トレーニングを行っても，「内容」と「オリジナリティ」という発表の内容に関する項目に対する学生の評価力を上げることができなかった可能性がある。今後，この点を見直し改善した評価者トレーニングを行うことで，「内容」と「オリジナリティ」の項目に対する学生の評価力に評価者トレーニングが及ぼす影響を再度検証する必要がある。また，これらの評価項目に対して何が影響を及ぼすのかを探求する必要がある。

3. 本研究で明らかになった評価者トレーニングの改善点である，評価項目，ルーブリックに書かれた判断基準の文言の明確さ，評価者トレーニングの方法，実施回数，および実施する時期を見直し，手順を簡易にした，授業内で実施可能な評価者トレーニングの方法についてさらに研究を進め，トレーニングを継続的に実施することで，学生の評価力を高められるかどうかを検証する必要がある。

評価を行う直前に一度に全評価項目の評価者トレーニングを行うには，多くの時間を要し，また学生にとっても一度に覚えるものが多く，負担が大きいものと思われた。しかし，評価項目別に評価者トレーニングを行う方法は，１回のトレーニングの所要時間を減らすことはできたが，結局各評価項目に対して１回ずつしか評価者トレーニングを行うことができず，結果として全評価項目の評価力を上げることができなかった。この点については，実施方法の見直す必要があり，今後の課題としたい。

今後，ルーブリックの文言を見直し，改善した評価項目別評価者トレーニングを行うことで，上記の項目に対する学生の評価力に及ぼす影響を再度検証する必要がある。この点についても，今後の課題としたい。

4. 評価者トレーニングが，学生のプレゼンテーションの指導の一つとして活用できるようにするために，音読練習やシャドーイングなどの音声トレーニングも併せることや，クラスメイトの発表の相互評価をするだけでなく，併せて自分のプレゼンテーションの振り返りとして，自己評価をさせることで，学生のプレゼンテーション力を向上させることができるかどうかを検証したいと考えている。

5. 本研究では，筆者が担当した大学１年生の１年間の成績だけを分析対象としたが，それから先の学生のプレゼンテーションの成績の推移も観察することで，評価者トレーニングの指導効果の可能性を確認したいと考えている。同じ学生を継続して担当し指導していくことは，制度上難しいことではあるが，できれば継続して担当する機会を持ちたいと考えている。具体的には，「プロジェクト発信型英語プログラム」におい

て，評価者トレーニングを導入し，その指導効果を確認するか，あるいは，小規模の大学において，自分が担当する学生を継続に担当することで，その指導効果を確認するかのいずれかであろう。また，大規模の大学では，全員が同じ基準で同じ方法で同じ回数評価者トレーニングを共通して行う方法が考えられる。

6. 「プロジェクト発信型プログラム」では，クラス分けを英語力に基づいて行っていないため，本研究の参加者の英語力は，TOEIC IP テスト平均 464 点と，一般的には英語力が高いとは言えない。また，本研究では，学生の英語力を測るのに，TOEIC IP テストの総合得点を用いたが，英語力の指標としては，ビジネスにおけるコミュニケーション能力を問う TOEIC テストを用いるのではなく，むしろ一般的な英語力や，授業で扱う内容に即した英語力を測れるテストを用いるべきであろう。

7.本研究では，他の担当教員との共同研究という形にできなかったこと，また，筆者担当のクラス数は，大学から割り当てられたもので，担当クラス数を増やすことができなかったことから，データ数が少ないということは本研究の問題点である。このような制約を受けた本研究の結果から，学生のプレゼンテーション力は，評価者としての学生の評価力には影響を及ぼさないという可能性が示されたと思われる。

今回得られた結果を再現できるかどうかを検証するために，今後，この研究を共同研究とし，データ数を十分に確保した上で追試を行う必要がある。

8.「プロジェクト発信型英語プログラム」は，学科全体の統一授業のため、様々な制約があった。一つは，評価の対象となるプレゼンテーションが学生一人ひとり異なっていることがである。評価の対象となるプレゼンテーションの題材は，オリジナリティは入れたうえで，学生全員同じものとする。また，授業で学習した内容に関して，自分のリサーチした成果を含めたプレゼンテーションにするなどの条件設定が必要であろう。2 つ目としては，評価シートに書かれた評価項目や判定基準が定められていたことである。3 つ目としては，相互評価を実施する際，中間発表ではクラス内で行っていたが，最終発表ではプログラム全体としてオーディエンスとなる学生を入れ替えて相互評価を行っていたことから，中間発表で相互評価を行った学生と最終発表で相互評価を行った学生が異なるため，厳密な両発表での比較ができなかったことである。これらの点が本研究の問題点であったため，今後はこれらを改善して研究を行う必要がある。

本研究では，上述したような問題点や限界があるが，学生の評価傾向を知ることは，評価項目，判定基準，相互評価の実施方法の再検討する良い機会となり得る。今後も，

学生に相互評価をさせる意義を引き続き考えていきたい。

注

1. Taylor & Galaczi (2011)は，評価力には，評価者自身の様々な背景が絡み合って，評価項目や判定基準の解釈や判断の仕方だけでなく，評価が甘目になったりあるいは厳しめになったりといった評価の一貫性にも影響を及ぼすと指摘している。

参考文献

馬場哲生 (編著). (1997). 『英語スピーキング論—話す力の育成と評価を科学する』東京：河源社.

Brown, H. D. (1993). *Principles of language learning and teaching* (3rd ed.) N.J.: Prentice Hall Regents.

Brown, J. D. (Ed.). (1998). *New ways of classroom assessment*. Alexandria, VA.: Teachers of English to Speakers of Other Languages, Inc. (TESOL).

Cheng, W., & Warren, M. (2005). Peer assessment of language proficiency. *Language Testing*, *22*(1), 93–121.

De Grez, L. (2010). Peer assessment of oral presentation skills. *Procedia Social and Behavioral Sciences*, *2*, 1776–1780.

Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing*, *11*, 125–144.

Falchikov, N, & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research*, *70*(3), 287-322.

Freeman, M. (1995). Peer assessment by groups of group work. *Assessment and Evaluation in Higher Education*, *20*(3), 289–300.

藤原康弘・大西仁・加藤浩 (2007a). 「公正な相互評価のための評価支援システムの開発と評価—学習成果物を相互評価する場合に評価者の選択で生じる「お互い様効果」ー『日本教育工学会論文誌』*31*(2), 125-134.

藤原康弘・大西仁・加藤浩 (2007b). 「学習者間の相互評価に関する研究の動向と展望」『メディア教育研究』*4*, 77-85.

深澤真 (2009).「スピーチにおける生徒相互評価の妥当性-項目応答理論を用いて」『STEP BULLETIN』, *21*, 31-47.

Fukazawa, M. (2010). Validity of peer assessment of speech performance. *Annual Review of English Language Education in Japan, 21*, 181–190.

Fulcher, G. (2003). *Testing Second Language Speaking*. Essex : Peason Education.

Goh, C., & Burns, A. (2012). *Teaching speaking*. Cambridge, England: Cambridge University Press.

Hanarahan, J., S., & Issacs, G. (2001). Assessing self- and peer-assessment: the students' views. *Higher Education Research & Development, 20*(1), 53–70.

Hirai, A., Ito, N., & O'ki, T. (2011). Applicability of peer assessment for classroom oral performance. *Japan Language Testing Association Journal, 14*, 41-59.

Hughes, A. (1989). *Testing for language teachers*. Cambridge, England: Cambridge University

Press.

Hughes, I. E. & Large, B. J. (1993). Staff and peer-group assessment of oral communication skills. *Studies in Higher Education, 18*(3), 379-385.

石川祥一・西田正・斉田智里 (編). (2011). 『テスティングと評価 －4 技能の測定から大学入試まで』東京：大修館書店.

笠巻知子 (2016).「スピーキング・パフォーマンス評価 -学生相互評価の信頼性- 」『言語と文化』*10*, 1–10. 京都外国語大学.

笠巻知子 (2018). 「学生の「プレゼンテーション力」は，評価者としての学生の評価力に影響を及ぼすか？」*Language Education & Technology, 55*, 97-122.

Luoma, S. (2004). *Assessing speaking*. Cambridge, England: Cambridge University Press.

McNamara, T. F. (1996). *Measuring second language performance*. London, England: Longman.

McNamara, T. F. (2000). *Language testing*. Oxford, England: Oxford University Press.

三木訓子・笠巻知子 (2017). 「Project-based English Program における学生による相互評価及びその可視化の試み」『立命館高等教育研究 』*17*, 165–181. 立命館大学.

望月昭彦・深澤真・印南洋・小泉理恵 (編著) (2015). 『英語 4 技能評価の理論と実践』東京：大修館書店

Nakamura, Y. (2002). Teacher assessment and peer assessment in practice. *Educational Studies, 44*, 203–215.

Norman, E. G. (1998). *Assessment of student achievement* (6th ed.) . Boston, MA:Allyn and Bacon.

Oi, S. Y. (2012). A pilot study of self-evaluation and peer evaluation. *Selected Papers of the 17th Conference of Pan-Pacific Association of Applied Linguistics*, 1-11.

岡秀夫 (編集). (1984). 『英語のスピーキング』東京：大修館書店.

岡田靖子 (2017). 「英語スピーキング・パフォーマンスにおけるピア評価の匿名化」『言語教育研究』*9*, 69-86. 清泉女子大学.

Okuda, R., & Otsu, R. (2010). Peer assessment for speeches as an aid to teacher grading. *The Language Teacher, 34* (4), 41–47.

Otoshi, J. & Heffernan, N. (2007). An analysis of peer assessment in EFL college oral presentation classrooms. *The Language Teacher, 31* (11), 3-8.

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

Saito, H. (2008). EFL classroom peer assessment : Training effects on rating and commenting *Language Testing, 25* (4), 553-581.

Sato, T. (2011). The contribution of test-taker's speech content to scores on an English oral

proficiency test. *Language Testing, 29* (2), 223-241.

Shimura, M. (2006). Peer and instructor assessment of oral presentations in Japanese University EFL classrooms : A pilot study. *Waseda Global Forum 3*, 99-107.

静哲人・竹内理・吉澤清美 (2002). 『外国語教育リサーチとテスティングの基礎概念』関西大学出版部.

菅沼洋子 (2013).「自己評価と他己評価を利用した自律的英語学習の探求 ー高校生によるスピーチを対象としてー」. 『STEP BULLETIN』, *25*, 168-185.

鈴木秀明 (2005). 「短時間の評価トレーニングが教師の発話評価に及ぼす効果」言語化学研究 *11*, 77-94.

鈴木佑治 (2012). 『グローバル社会を生きるための英語授業：立命館大学　生命科学部・薬学部・生命科学研究科　プロジェクト発信型英語プログラム』 東京 : 創英社/三省堂書店.

Tanaka, M. (2017). Examining personality bias in peer assessment of EFL oral presentation : A preliminary study. *JALT Journal, 39* (2), 183-196.

Taylor, L. & Galaczi, E. (2011). Scoring validity. In L. Taylor (Ed.), *Examining speaking* (pp.171-233). Cambridge, England: Cambridge University Press.

Underhill, N. (1987). *Testing spoken Language*. Cambridge, England: Cambridge University Press.

Ur, P. (2012). *A course in English language teaching*. Cambridge, England: Cambridge University Press.

Weigle, S.C. (1994). Effects of training on raters of ESL composition. *Language Testing, 11*, 197–223.

山西博之 (2004).「高校生の自由英作文評価はどのように評価されているのか —分析的評価尺度と総合的評価尺度の比較を通しての検討—」 *JALT Journal, 26*, 189–205.

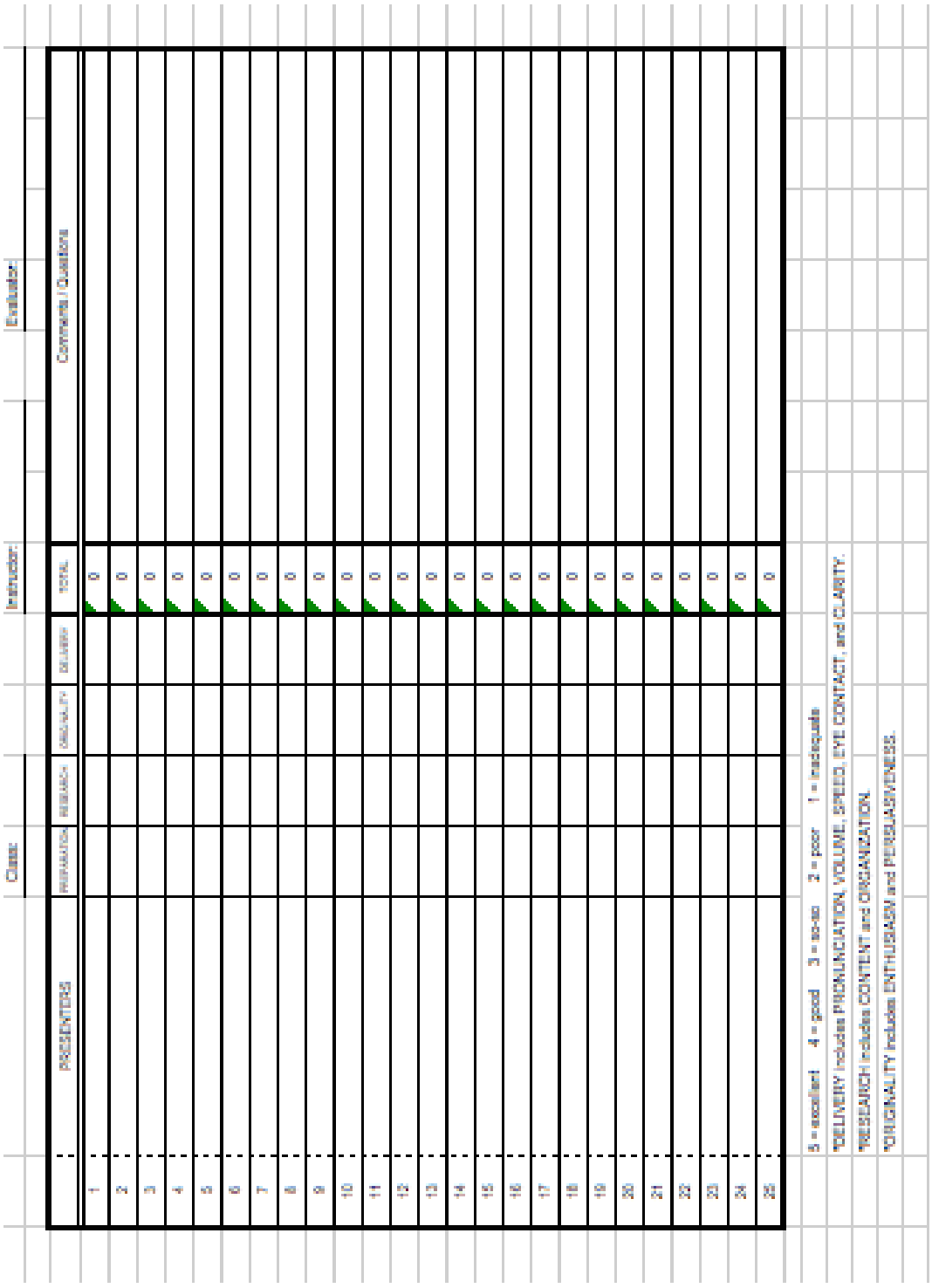
Yang, M., Badger, R., & Yu, Z. (2006). A comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of Second Language Writing, 15*, 179-200.

Appendix

資料 1：評価シート

評価項目：準備、リサーチ、オリジナリティ、発表の仕方

評価基準：excellent = 5、good = 4、so so = 3、poor = 2、inadequate = 1



資料２　評価シート

評価項目：内容，オリジナリティ，アイコンタクト，スライド，発音，ポーズ

	PRESENTERS	content	originality	eye-contact	slide	pronunciation	speed	TOTAL	Comments / Questions
1								0	
2								0	
3								0	
4								0	
5								0	
6								0	
7								0	
8								0	
9								0	
10								0	
11								0	
12								0	
13								0	
14								0	
15								0	
16								0	
17								0	
18								0	
19								0	
20								0	
21								0	
22								0	
23								0	
24								0	
25								0	