JASLA Research Report
Vol.2017, No.1 (2017-5)

学生の相互評価における評価タイミング取得の必要性

堀越泉*、田村恭久**

*上智大学大学院理工学研究科
**上智大学理工学部

Timing Acquisition of Students' Peer Assessment

Izumi HORIKOSHI*, Yasuhisa TAMURA**

* Graduate School of Science and Technology, Sophia University
** Dept. Information and Communication Sciences, Sophia University

* izumihorikoshi@eagle.sophia.ac.jp

概要：本稿では、学生による相互評価において評価タイミングを含めて信頼性・妥当性を分析する必要性を検討した。一般に、相互評価の信頼性・妥当性を検証には評点の一致度が用いられる。本研究では、評価項目の評点と評価時刻の自動取得機能を持つ相互評価ツールを開発し、被験者 5 名に対しプレゼンテーションの相互評価を実施し、19 項目の評点と評価時刻を取得・分析した。この結果、評点をもとにした従来手法では信頼性・妥当性が中程度と判定されたが、評価時刻は必ずしも一致していないという結果を得た。また、複数項目を短時間で記入するなどの行動が観察された。

Abstract: In this paper, the authors focused on students' peer assessments. The purpose of this study was to discusses the need to consider temporal behavior in their assessment process. In general, verification of reliability or validity of peer assessment has been carried out by analyzing their score. In this research, the authors developed a peer assessment tool with the function to detect students' temporal behavior in peer assessment, and analyzed the assessment process. The result of the conventional method of verification with score results showed that students' peer assessments was moderately reliable and valid. However, students did not necessarily evaluate in same timings. The authors also found a characteristic behavior that student scores many items in short time period.

キーワード： Learning Analytics、相互評価、信頼性、妥当性、評価タイミング
Keywords: Learning Analytics, peer assessment, reliability, validity, temporality

1. はじめに

1.1 背景

近年、学生にプレゼンテーションをさせ、学生同士で相互評価させるという形式の授業が増えてきている。筆者らの担当している授業でも、様々な学年の授業に於いて、プレゼンテーションと相互評価の授業に取り入れている。学生に相互評価させる目的は授業者によって様々であると考えられるが、筆者らは主に以下の2つの目的を持って授業に相互評価を取り入れている。

- 目的 A：学生自身が発表者と評価者を経験することで、評価基準(=到達目標)をより理解し、プレゼンテーション技術が向上することを期待して
- 目的 B：大人数授業でプレゼンテーションを実施し教員が全発表を見られない際に成績に利用するため

学習場面における相互評価の一般的な利点については、深澤(2010)が文献を整理している。

- 学習目標をはっきり認識させる(Luoma, 2004)
- 学習者の動機付けを高める(Orsmond,1996)
- 評価における教員負担の削減(Brown,1998)

筆者らの挙げた目的1、2はそれぞれ Luoma(2004)、

Brown(1998)に相当すると言える。

1.2　着目した問題

一方で、学生による相互評価には信頼性・妥当性の問題がある。深澤(2010)は前出の報告の中で、相互評価を総括的評価（成績）の一部として用いる場合妥当性を検証する必要がある、と指摘した上で、学生による相互評価の信頼性・妥当性に関する関連研究の整理も行なっている。深澤の報告によると、関連研究は大きく分けて、学生による相互評価は教員評価と比較しうる（信頼できる）という報告と、信頼性・妥当性には疑問があるという報告に分かれる。前者の例として、教員評価と相互評価の相関を分析し妥当性を検証して、高い相関が見られた３つの研究を挙げている。それぞれの研究における教員評価と相互評価の間の相関係数は以下の通りである。r=.68-.80 (Miller, 1996)、r=.83 (Hughes, 1993)、r=.89 (Stefani, 1994)。これに対して、学生による相互評価は信頼性・妥当性には疑問があるという報告の例として、「平均値は教員評価のより相互評価の方が高く、標準偏差は教員評価のより相互評価の方が小さい傾向にある」と報告した Stefani(1994)や、「平均値は教員評価のより相互評価の方が高く、標準偏差は教員評価のより相互評価の方が小さい傾向にある」と報告した Freeman (1995)を挙げている。

ここで我々が着目したのは、いずれの関連研究も、基本的に相互評価の評点を元に信頼性・妥当性を議論している、ということである。信頼性は学生同士の評価の一致度、妥当性は教員による評価と学生による評価の一致度で議論されていることが多い。しかし、各評価項目を評価したタイミングに着目した時、「全く異なる時点に偶然近い評点をつける」という場合があるのではないか、と考えた。もしこのようなことが起こっているのであれば、評価したタイミングが異なっていても評点が一致していれば「評価が一致している」とみなしていることになる。また、教育の一環としての相互評価の場合、評点からだけでは「適切に評価する能力」を伸ばすために学生にどのように指導したら良いか検討する情報が少ないという問題もある。

1.3　本稿の目的

本研究では相互評価における各評価項目の評価タイミングに着目する。本研究の効果として、基本的に評点を元に信頼性・妥当性を議論していた従来手法では見えていなかった課題を可視化し、信頼性・妥当性の保証のための議論や、信頼性・妥当性の向上のための指導に新しい視点を提供できると期待する。

本稿では、研究の予備調査として、以下の３点を目的とした。

- **目的１　実現可能性の検討**：相互評価の評価タイミングを取得するツールを開発し、予備実験を行う
- **目的２　研究意義の検討**：予備実験の結果より、評点を用いた従来手法による妥当性・信頼性の検証結果と、評価タイミングを用いた提案手法によって明らかになることを比較
- **目的３　仮説の詳細化**：予備実験のデータを分析・可視化し、仮説を詳細化して本実験設計に繋げる

2. 方法

2.1　評価タイミング取得ツールの開発

筆者らの授業では従来から相互評価を行なっており、ここでは Google 社のオンラインフォーム（以下 Google フォーム）を用いてきた。Google フォームは容易にアンケートフォームが作成可能であり、受講者に対しオンラインで配布し、回答を自動でスプレッドシートに集約することができる。さらに、提出時刻がタイムスタンプつきで記録される。ただし、この時刻はフォーム上の全ての評価項目に回答し終えて、最後に送信ボタンを押した時点の時刻である。

本研究では各評価項目の評価時刻を分析対象とするため、この粒度でログ取得が可能な相互評価用ツールを新たに開発した。このツールはオンラインフォームであり、HTML、JavaScript、PHP で実装した。このフォームは評価項目のリストと評点に対応したラジオボタン（1-3 または 1-5）、送信ボタンからなる。

今回の相互評価用オンラインフォームでは、評価を終えて送信ボタンを押した時点の評価時刻及び評点だけでなく、送信ボタンを押す前の各評価項目の評価時

刻及び評点も記録する。評価者は送信ボタンを押すまで、何度でも評価を変更可能であり、変更を行うたびにログが記録される。ログ項目中の「submit type」とは送信ボタンを押す前の評価（途中評価）か送信ボタンを押した時の評価（最終評価）かの区別である。ラジオボタンの選択が変更されるか、送信ボタンが押されるとログが Learning Record Store (LRS)に送られる。取得した主なログ項目は以下の通りである。

- 評価日時
- 評価者学生番号（記入者）
- 発表者学生番号（被評価者）
- 評価項目番号
- 評点
- 「submit type」（途中評価／最終評価）

2.2 実験及び分析設計

本稿の目的１・２のため、実験及び分析を設計した。

2.2.1 実験対象科目

開発した相互評価の評価タイミング取得するツールを用いた予備実験を実施し、実現可能性の検討を行うため（目的１）、授業形式の異なる2つの科目において実験を設定した（表 1）。どちらも上智大学の開講科目であり、筆者らが教員及びティーチングアシスタント（TA）を担当している。

ゼミナールⅡは理工学部3年生を対象とした科目で、学生4人のゼミナール形式の授業であり、毎回 30 分のプレゼンテーションが課される。実験対象回には2名の学生がプレゼンテーションを行なった。システムコンサルティングは約 80 名が履修する全学共通科目で、学期末にグループ学習の調査発表の５分のプレゼンテーションを行なった回を対象とした。

ゼミナール Ⅱ は被験者数が少ないが、プレゼンテーション時間が長いため評価過程のログが多く取得できること、また TA も評価に加わっているため、TA と学生の評価を比較する目的で対象科目に設定した。一方、システムコンサルティングはプレゼンテーション時間が短く、また TA は評価に加わっていないため TA と学生の評価は比較できないが、被験者数が多いため、大規模実験の実現可能性の検討のため対象科目に設定した。また、2科目でプレゼンテーション時間が大きく異なるため、この影響の分析も目的とした。

表 1　実験対象科目、単元、実施日および被験者数

対象科目	実施日	単元 ID	被験者数	授業形式
ゼミナールⅡ（SE）	2016/10/21	SE 1021	学生: 4 TA:1	プレゼンテーション 30 分 評価者： 学生 ＋TA 評価項目数*: 19
システムコンサルティング（LE）	2016/11/15	LE 1115	学生: 83 TA: -	プレゼンテーション５分 評価者： 学生のみ 評価項目数*: 5

*注：具体的な評価項目については付録 A に添付

2.2.2 分析

予備実験のデータを用いて、評点を用いた従来手法による妥当性・信頼性の検証結果と、評価タイミングを用いた提案手法によって明らかになることを比較し、研究意義の検討を行うため（目的２）、分析を設計した。

学生による相互評価の平均値を評価に用いる場合、評点を用いて妥当性・信頼性の検証を行う場合、以下のような分析を行うことが一般的である（従来手法）。

- 信頼性：
  - 学生間での各評価項目の評点の一致係数を算出
- 妥当性：
  - 評価の妥当性
    - 学生平均と教員等の評価の一致係数を算出
  - 評価の厳しさの妥当性
    - 学生平均と教員等の評価の平均値の差を検定
      （平均値に有意差がなければ妥当）

今回は、定量的分析として上記の⑴評点を用いた従来手法による信頼性・妥当性の検証、⑵評点により明

らかになる評価の傾向と評価タイミングにより明らかになる評価の傾向の比較、(3)その他価タイミングにより明らかになる特徴的な評価行動の集計、定性的分析として(4)評価タイミングの可視化、(5) TA と学生の評価タイミングの比較を行なった。実験対象科目の授業形式の特性から、ゼミナールII（SE）に対しては分析(1), (2)及び(4), (5)を、システムコンサルティング（LE）に対しては分析(3), (4)を行なった。

3. 結果

3.1　従来手法による信頼性・妥当性の検証

ゼミナールIIの相互評価の評点に対し、2.2.2 項に示した手法を用いて信頼性・妥当性の検証を行なったところ、表 2 のような結果となった。

表 2　評点を用いた従来手法による信頼性・妥当性

発表者	評価者数	学生間	学生平均-TA 間	
		ケンドールの一致係数	ケンドールの一致係数	t 検定
S3	4	W = .669	W = .500	t = 0.143 n.s.
S4	4	W = .767	W = .500	t = −0.343 n.s.

学生による相互評価の平均値を評価に用いる場合、信頼性については、学生評価間の一致係数が 0.66 及び 0.767 であったので中程度～かなりの信頼性であると言える。また妥当性については、学生評価の平均と TA 評価間の一致係数が 0.500 及び 0.500 であったので中程度の妥当性であると言える。さらに、評価の厳しさの妥当性についても、学生評価の平均と TA 評価間の平均値に有意差が見られなかったため、妥当であると言える。

従って、ゼミナール II の相互評価は、従来手法では信頼性・妥当性共に中程度であると言える。

3.2　評点・評価タイミングの傾向の比較

ゼミナール II について、評点により明らかになる評価の傾向と評価タイミングにより明らかになる評価の傾向を比較するため、各評価項目について評価タイミングと評点の分布を可視化した（図 1)。2名の学生がプレゼンテーションを行なったため2回分をプロットした。横軸は評価項目番号、縦軸はプレゼンテーション開始時刻からの経過時間(秒)である。

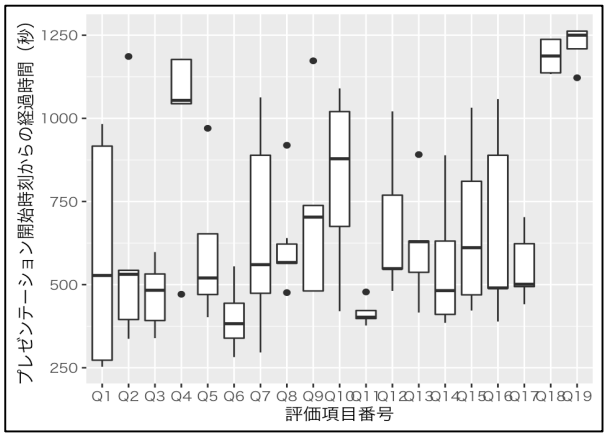
評価タイミングのばらつきが大きい項目、小さい項目はそれぞれ以下の通りである。

ばらつきが大きい項目

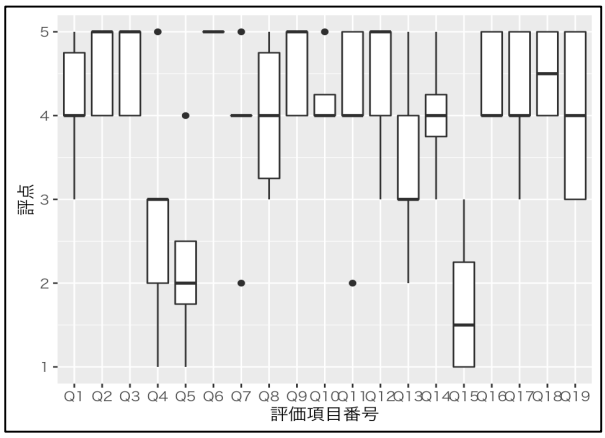
- Q1：タイトルは適切か？
- Q7：内容を理解していたか？
- Q15：指示棒の使用は適切だったか

毎回ばらつきが小さい項目

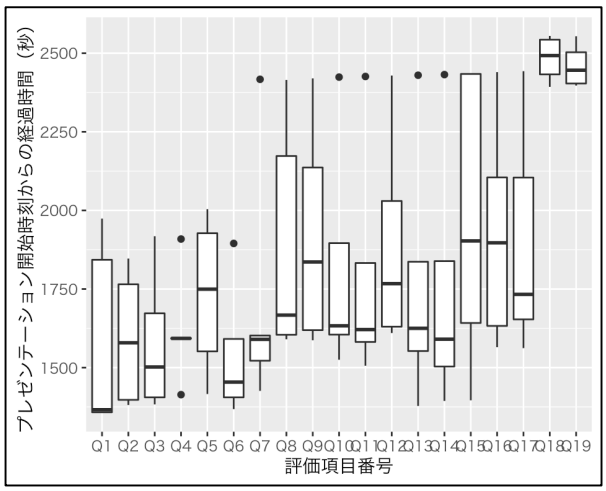
- Q18：質問と回答は合致していたか？
- Q19：結論から答えていたか？



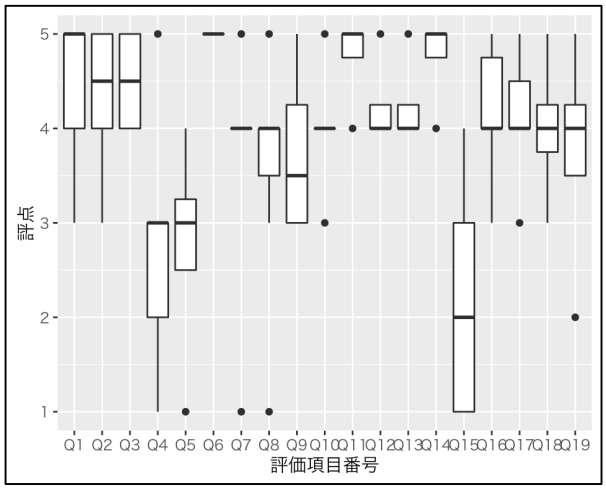
(a-1) 評価１回目（評価タイミング）



(b-1) 評価１回目（評点）



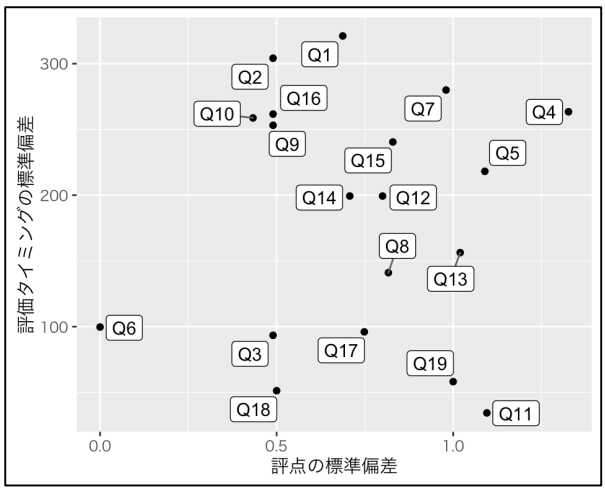
(a-2) 評価 2 回目（評価タイミング）



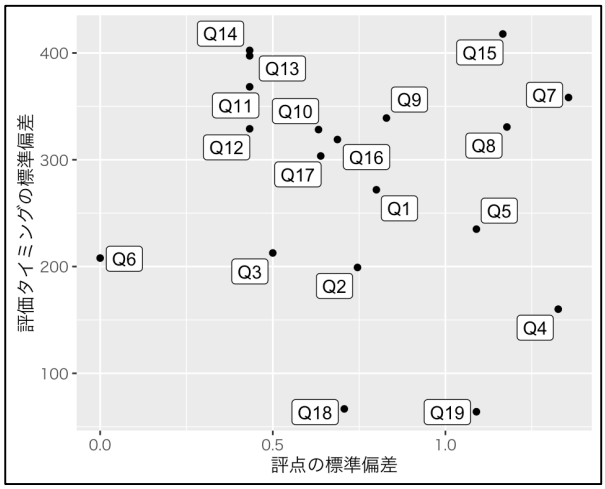
(b-2) 評価 2 回目（評点）

図 1　評価タイミングと評点の分布

評点と評価タイミングの傾向を比較すると、評価 2 回目の Q1～14 のように評点のばらつきは小さいにも関わらず、評価時刻のばらつきは大きい項目が見られた。評点と評価タイミングのばらつきの関係を可視化するため、評点の標準偏差と評価タイミングの標準偏差の散布図を作成すると図 2 のようになった。縦軸が評価タイミングの標準偏差、横軸が評点の標準偏差である。定量的に評価するため、評点の標準偏差と評価タイミングの標準偏差の相関分析を行なったところ、評価 1 回目・2 回目共に相関は見られなかった（1 回目：r = .0181, n.s., 2 回目：r = - . 112, n.s.）つまり、評点のばらつきが小さくても、評価タイミングのばらつきが小さいとは限らず、評点と評価タイミングの間は無相関であることが分かった。



(a) 評価 1 回目

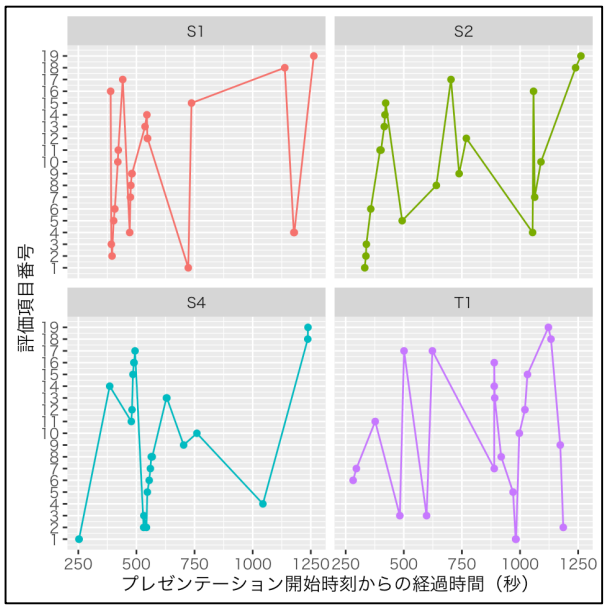


(b) 評価 2 回目

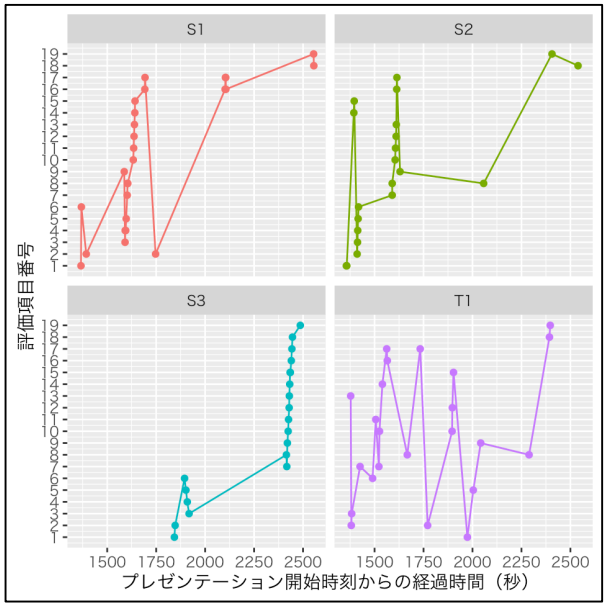
図 2　評点の標準偏差と
評価タイミングの標準偏差の関係

3.3　評価タイミングの可視化

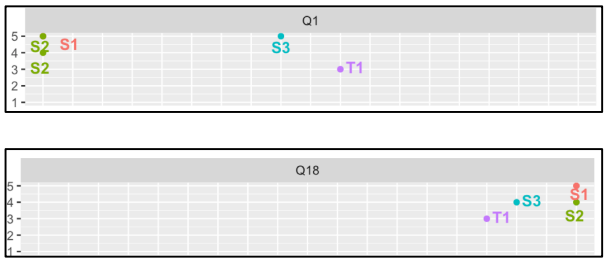
ゼミナール II 及びシステムコンサルティングにおいて、開発したツールによって取得した評価タイミングをプロットすると図 3、図 4 のようになった。横軸はプレゼンテーション開始時刻からの経過時間（秒）、縦軸は評価項目番号である。ゼミナール II では2名の学生がプレゼンテーションを行なったため2回分をプロットした。システムコンサルティングでは全 12 グループがプレゼンテーションを行なったが、このうち最初のグループのみをプロットした。図 3、図 4 中の英数字は評価者番号であり、S から始まる評価者は学生、T から始まる評価者は TA である。



(a) 評価１回目（発表者 S3）



(b) ゼミナールⅡ 評価２回目（発表者 S4）



横軸：プレゼンテーション開始時刻からの経過時間、縦軸：評点

(c) 評価２回目　抜粋（Q1，Q18）

図 3 評価タイミング（ゼミナールⅡ）

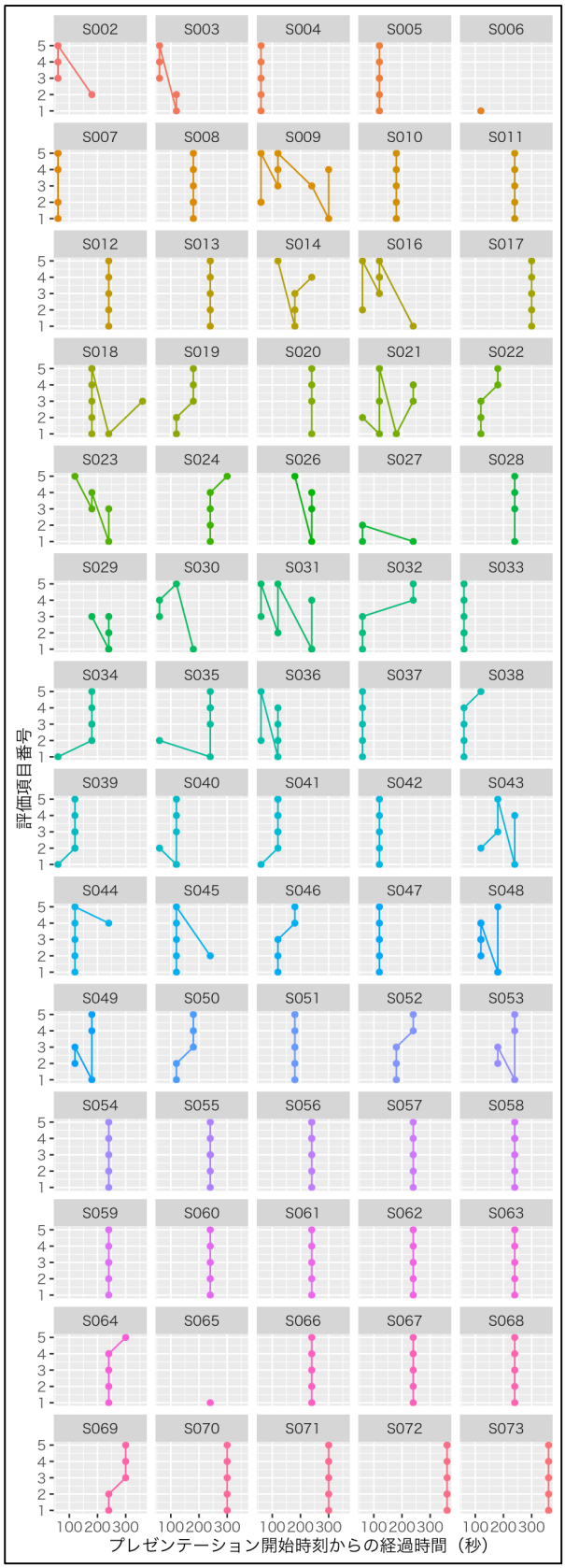


図 4 評価タイミング（システムコンサルティング，発表時間外につけられた評価は除外）

まず、ゼミナール II のログについて、学生による評価(S1〜S4)と TA による評価(T1)を比較すると、特に2回目の評価において、学生は複数項目を短時間に評価項目番号順に記入していると見られる行動が見られ、一方の TA はプレゼンテーション時間全体にわたって、評価項目番号と関係なく評価する行動が見られた。また、評価項目 Q1(「タイトルは適切か？」)と Q18(「質問と回答は合致していたか？」)に着目すると、Q1 は学生と TA で評価タイミングが分かれているのに対し、Q18 は学生も TA も同じようなタイミングで評価していることが分かる。

システムコンサルティングのログでは、発表時間外に評価した学生が 13. 5%存在した。この学生を除いてプロットしたグラフが図 4 である。図 4 より、複数項目を短時間で評価項目番号順記入する行動が多く見られる。発表時間内に評価した学生のうち、評価にかけた時間が 2 分未満であった学生は 72.4 %を占めた。

4. 考察

4.1 定性的議論

3.3 節に示したように、評価者ごとに評価タイミングをプロットして可視化すると、複数項目を短時間で記入するなどの行動が見られた。特に、評価するプレゼンテーションの時間が短く、大人数教室で行なったシステムコンサルティングの相互評価で顕著であったことから、環境や条件によって真剣に考えながら相互評価を行う難しさが変化するのではないか、という新しい仮説が得られた。

また、評価タイミングのばらつきが特に大きい項目がいくつか見られたが、この原因はいくつかのパターンに分けられる可能性が示唆された。例えば、Q1(「タイトルは適切か」)は TA と学習者の評価タイミングが異なり、学生は評価の最初にタイトルについて評価していたのに対し、TA は中盤以降に評価していた。これは、TA はある程度プレゼンテーションが進んだ時点で発表内容とタイトルの整合性を考えて評価しているのに対し、学生は評価項目番号に従って最初に評価していると考えられ、着目すべきポイントを指導すべき項目であると言える。一方、Q15(「指示棒の使用は適切だったか」)のばらつきも大きいが、この項目は評価に適したタイミングが複数存在すると考えられる。このような項目も存在するため、評価タイミングのばらつきが大きいと信頼性が低く、また評価タイミングが教員と異なっていると妥当性が低いと一概には言えないということが明らかになった。

4.2 定量的議論

3.1 節に示したように、評点をもとにした従来手法では信頼性・妥当性が中程度と判定されたが、3.2 に示したように評点の標準偏差と評価タイミングの標準偏差の間に相関は見られなかった。これは、評点のばらつきが小さくても、評価タイミングのばらつきが小さいとは限らないことを示している。つまり、評価タイミングの分析から得られる情報は評点の分析から得られる情報では代替できず、学生の相互評価において評価タイミングを取得する必要性を示唆している。また、「全く異なる時点に偶然近い評点をつけるという場合があるのではないか」という仮説は示されたことになる。

5. 結論及び今後の課題

本稿の目的は、(1)相互評価の評価タイミングを取得するツールを開発し、予備実験を行う（実現可能性の検討）、(2)予備実験の結果より、評点を用いた従来手法による妥当性・信頼性の検証結果と、評価タイミングを用いた提案手法によって明らかになることを比較（研究意義の検討）、(1)予備実験のデータを分析・可視化し、仮説を詳細化して本実験設計に繋げる（仮説の詳細化）の３点であった。

開発した相互評価の評価タイミング取得するツールを用いた予備実験を実施し、大人数授業であるシステムコンサルティングにおいてもログを取得できたことから、目的１は達成できたと言える。

目的２については、評点の標準偏差と評価タイミングの標準偏差の間に相関は見られなかったという結果から、評価タイミングの分析から得られる情報は評点の分析から得られる情報では代替できず、学生の相互評価において評価タイミングを取得する必要性が示唆される、という結論に至った。また、今回は定性的な議論に留まったものの、評価タイミングの分析が相互評価の状況把握や課題把握に寄与する可能性を得られた。

今後は、今回の予備実験の結果より得た以下の仮説

をより詳細化し、明らかにしていきたい。特に、仮説２・３のように、授業設計改善や指導を行った際の変化を可視化・分析し、相互評価の形成的評価に寄与する知見を得たい。

- 仮説１：短時間に複数項目を記入している行動は「評価項目番号順」に記入しており、考真剣に考えながら相互評価を行なっていないことが現れた行動である
- 仮説２：授業人数やプレゼンテーション時間の長さ、評価項目数によって相互評価を行う難しさが変化する
- 仮説３：評価項目の文言を改善したり、評価の際に着目すべきポイントを指導することにより、評価タイミングが一致するようになる
- 仮説４：評価タイミングのばらつきが大きい項目の中には、評価に適したタイミングが複数存在するものがある

謝辞

本研究はJSPS科研費26282059（基盤研究(B)）の助成を受けたものです。

参考文献

Brown, J. D. (1998). *New Ways of Classroom Assessment. New Ways in TESOL Series II. Innovative Classroom Techniques*. TESOL, 1600 Cameron Street, Suite 300, Alexandria, VA 22314-2751

Freeman, M. (1995). Peer assessment by groups of group work. *Assessment & Evaluation in Higher Education*, *20*(3), 289-300.

深澤真. (2010). スピーチにおける生徒相互評価の妥当性. *ARELE: annual review of English language education in Japan*, 21, 181-190. (in Japanese)

Hughes, I. E., & Large, B. J. (1993). Staff and peer-group assessment of oral communication skills. *Studies in Higher Education*, *18*(3), 379-385.

Luoma, S.(2004). *Assessing speaking*. U.K.: Cambridge University Press.

Miller, L., & Ng, R. (1996). Autonomy in the classroom: Peer assessment. *Taking control: Autonomy in language learning*, 133-146.

Orsmond, P., Merry, S., & Reiling, K. (1996). The importance of marking criteria in the use of peer assessment. *Assessment & Evaluation in Higher Education*, *21*(3), 239-250.

Stefani, L. A. (1994). Peer, self and tutor assessment: relative reliabilities. *Studies in Higher Education*, *19*(1), 69-75.

付録Ａ

評価項目（ゼミナールII:SE）

- Q1: タイトルは適切か？
- Q2: 本文の行数は適切か？
- Q3: 字数を減らす工夫をしていたか？
- Q4: 数字の表をグラフ化していたか？
- Q5: 色は見やすく使っていたか？
- Q6: スライド番号を付けたか？
- Q7: 内容を理解していたか？
- Q8: 論理的に構成されていたか？
- Q9: 不明語彙の意味を調べたか？
- Q10: 発声の音量・速度は適切だったか？
- Q11: 聴衆に向かって話していたか？
- Q12: 長い沈黙がなかったか？
- Q13: アイコンタクトをとっていたか？
- Q14: 立ち位置は適切だったか？
- Q15: 指示棒の使用は適切だったか？
- Q16: スライド送りはスムーズだったか？
- Q17: 次スライドへのつなぎは適切だったか？
- Q18: 質問と回答は合致していたか？
- Q19: 結論から答えていたか？

評価項目（システムコンサルティング：LE）

- Q1: 説明された考えが論理的かつ明確であったか
- Q2: 説明のスライドはわかりやすく、説得力が高く構成されていたか
- Q3: 説明担当者は内容を理解して説明していたか
- Q4: 説明担当者の指示棒の使用、ジェスチャー、アイコンタクトは良好だったか
- Q5: 説明担当者の滑舌や声量は良好だったか