JSiSE Research Report
vol.35,no.3(2020-9)

相互評価中の特徴的な評価行動の指標に対する連続評価の影響分析

堀越 泉[*1]，田村 恭久[*2*]
*1 上智大学大学院 理工学研究科，*2 上智大学 理工学部

Analysis of The Effects of Continuous Evaluation on the Indicators of Characteristic Evaluation Behavior during Mutual Evaluation

Izumi HORIKOSHI[*1], Yasuhisa TAMURA[*2]
*1Graduate School of Science and Technology, Sophia University,
*2Faculty of Science and Technology, Sophia University

本稿では，相互評価中の評価行動が連続評価の際に受ける影響を分析した．以前の報告では，特徴的な相互評価行動を定量的に表現できる指標として，評価にかけた時間，クリック回数，時刻や評点の平均・標準偏差などを提案した．本稿では，これらの指標が連続評価の影響を受けやすいかを分析した．その結果，グラフの形状に現れるレベルではどの指標も連続評価に伴う単純な減少・増加は見られなかった．一方で，評価者間でのばらつきが大きいものと小さいものが見られた．特にクリック回数については，分散分析においても評価者間で差異が少ないことが明らかになった．

キーワード: 相互評価，評価行動，Learning Analytics

1. はじめに

近年，学力観の変化等を背景として，学習者による相互評価が盛んになってきている(1)．学習場面における相互評価の一般的な利点として，深澤 (2)は学習者の動機付けを高める(3)，学習目標をはっきり認識させる(4)，評価における教員負担の削減(5)などを挙げている．

一方で，学習者同士による相互評価は，学習者が真面目に評価を行っているか，あるいは真面目に評価を行っていたとしてもその評価は妥当であるか，など懸念がある．この問題に対し，学習者による相互評価の信頼性・妥当性を分析した先行研究は多く存在する(2)．学習者による相互評価は教員評価と比較しうる（信頼できる）という文献(6),(7)がある一方，「教員評価と相互評価の間の相関は中程度であり，また良い発表を低く評価し，よくない発表を高く評価する傾向が見られた」というように，信頼性・妥当性には疑問があるという文献(8)もある．また，評価する学習者が，評価対象となっている学習者からも評価されると甘い評価を行う傾向があるという文献(9)もあり，相互評価を行う条件が評価結果に影響を与える可能性も指摘されている．

このように．相互評価の質の議論については多数の先行研究が存在する．これらの先行研究では，主に学習者同士の評点の一致度をもって信頼性を，教員による評点と学習者による評点の一致度をもって妥当性を議論するなど，基本的に相互評価の評点（結果）に着目している．

これに対し筆者らは，一人一人の学習者がいつ・どの評価項目を評価したかという評価プロセスに着目し，評価行動を分析してきた．そしてこれまでの結果として，評価にかける時間が長い学習者や短い学習者が存在すること，評価項目の並びと異なる順序で評価する学習者や，評価項目順に評価する学習者が存在するこ

となどを発見し，学習者によって評価行動は様々であるということを明らかにしてきた．

相互評価の質の議論において評価行動に着目した研究はほとんど存在しない一方で，社会調査の分野では，近年 Web を用いたアンケートの回答にかけた時間や中断などのログを分析し，調査の回答の質の検証に用いる研究が盛んになってきている[10]．そして例えば「アンケート回答の繰り返しに伴い，回答者の正確に答える動機づけが低下し，短時間に回答する行動が増加する」[11]などの知見が蓄積してきている．このため，相互評価の評価行動においても連続評価の実施により，評価に対する動機付けに影響を与える可能性があると考えられる．

既報の研究は一人一人の評価行動の可視化をもとにした定性的な議論であった．この手法では，個人の行動の特徴については詳細に議論できる．一方で，特徴間の関係やクラス全体の傾向に対する議論には不向きである．そこで本稿では，評価行動の特徴を定量的に議論する手法を検討する．これを実現するため，評価行動の特徴を定量的な指標（特徴量変数）として抽出し，その傾向や性質を明らかにすることを目的とする．

2. 方法

2.1 データ取得方法

相互評価中の各評価項目の評点と評価タイミングを取得するため，既報の研究でも使用してきた相互評価フォームを用いた．これはオンラインフォームであり，評価項目のリストと評点に対応したラジオボタン（1点－5点），送信ボタンを備える．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，評点の選択，変更，送信ボタンの押下をトリガーとして，履歴がサーバに送られる．取得した主なログ項目は，評価日時，評価者学生番号，被評価者学生番号，評価項目番号，評点である．

2.2 実験

上記のフォームを用いて実験を実施し，相互評価の履歴を取得した．被験者は上智大学開講の全学共通科目「情報リテラシー（情報検索）」受講者のうち，被験者として同意を得た者とした．実験対象回では，プレゼンテーション作成の演習を行い，その発表に対し相互評価を課した．プレゼンテーションは発表 10 分，質疑応答 4 分からなり，グループ A～F の計 6 グループが発表を行なった．実験実施日は 2017 年 7 月 10 日，被験者は 71 名であった．

2.3 抽出する特徴量変数

上記の方法で取得した相互評価中の各評価項目の評点と評価タイミングのデータから，以下の 6 種の特徴量変数を抽出した．

評価にかけた時間（Evaluation Time: ET):
評価者が最初に評価項目をクリックしてから提出ボタンを押す前に最後に評価項目をクリックするまでの経過時間（注:「フォームが表示されてから送信ボタンが押されるまでの経過時間」ではない）．

クリック回数(Click Count: CC):
評価者が評価を通じて評価項目のラジオボタンをクリックした合計回数．

評点の平均(Mean of the score: sM):
評価者が使用した評点の平均．

評点の標準偏差(Standard deviation of the score: sSD):
評価者が使用した評点の標準偏差．

評価時のタイムスタンプの平均(Mean of the evaluation time stamp: tM): 評価項目クリック時のプレゼンテーション開始時からの平均経過時間．

評価時のタイムスタンプの標準偏差(Standard deviation of the evaluation time stamp: tSD):
評価項目クリック時のプレゼンテーション開始時からの経過時間の標準偏差．

2.4 Research Question と検証方法

本稿では， 2 つの Research Question(RQ)を設定し，以下の方法で検証した．

RQ１：特徴的な評価行動は定量的に表現可能か

上で提案した 6 種の特徴量変数と関連する特徴的な評価行動を抽出し，その特徴を特徴量変数を用いて表現

RQ２:特徴量変数によって連続評価から受ける影響に差異があるか

連続評価を行なった際の特徴量変数の推移を可視化し，比較．連続評価の影響と個人差の影響を分析

3. 結果・考察

3.1 取得データ

前述の実験の結果，**表 1** に示す件数のデータが得られた．本稿で対象とする回にプレゼンテーションを行なったのはグループ A～F の 6 グループであったが，誤ってグループ G の評価を行なった評価者がいたため，**表 1** にはグループ G の欄も存在する．また，重複して評価を送信した学習者や途中まで評価して送信ボタンを押さなかった学習者などが存在したが，**表 1** では最終的に提出ボタンを押した学習者の重複を除いて評価件数とした．

表 1 評価対象グループ別評価件数

評価対象グループ	評価件数
グループ A	49
グループ B	52
グループ C	52
グループ D	52
グループ E	49
グループ F	51
グループ G	1

3.2 RQ１：特徴的な評価行動は定量的に表現可能か

RQ1 を検証するため，上で提案した 6 種の特徴量変数と関連する特徴的な評価行動を抽出し，その特徴を特徴量変数を用いて表現する．次頁**図 1** は，今回得られた特徴的な評価行動である．以下では**図 1** で取り上げた評価行動と特徴量変数を整理する．

評価にかけた時間（Evaluation Time: ET）– 図 1(a)

評価にかけた時間（ET)は，評価者が評価に時間をかけたかどうかを表現することができる． **図 1** (a-1)に取り上げた「時間をかけて評価を行った評価者」は ET=12.3(分)であるのに対し，(a-2)の「短時間で評価を行った評価者」は ET=0.73(分)であった．

クリック回数（Click Count: CC）– 図 1(b)

クリック回数(CC)は評価者が評価の変更を行ったかどうかを表現することができる．**図 1** (b-1)の評価者はいちどつけた評価の変更を行ったため評価項目数より多くクリックし，CC=21(回)であった．一方，(b-2)の評価者はほとんどの評価項目について変更を行わなかったため評価項目数とほぼ同じ回数のみクリックし，CC=18(回)であった．

評点の平均（Mean of the score: sM）– 図 1(c)

評点の平均(sM)は評価者の評点選択傾向を表現できる．**図 1** (c-1)の評価者は低得点を含む様々な評点を使用したために平均の評点は低くなり sM=3.63(点)であったのに対し，(c-2)の評価者は高得点の評点を多く使用したため平均の評点は高くなり sM=4.45(点)であった．ただし，評点の平均(sM)のみからでは，評価者が同一の評点ばかりを使用したかどうかまでは表現できない．

評点の標準偏差（Standard deviation of the score: sSD）– 図 1(d)

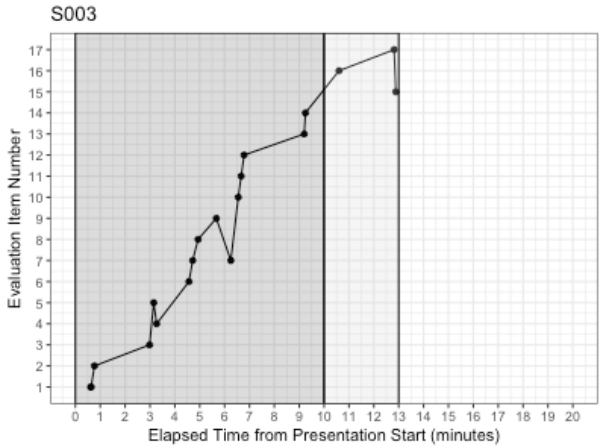
評点の標準偏差(sSD)は，評価者が同一の評点ばかりを使用したのか，様々な評点を使用したのかを表現できる．**図 1** (d-1)の評価者は様々な評点を使用し sSD=0.99(点)であったのに対し，(d-2)の評価者は特定の評点をばかり使用したため sSD=0.34(点)と(d-1)より小さい値であった．

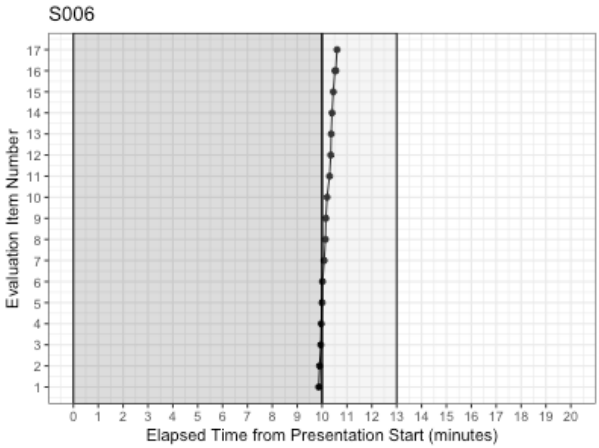
評価時のタイムスタンプの平均(Mean of the evaluation time stamp: tM) – 図 1(e)

評価時のタイムスタンプの平均(tM)は評価者がプレゼンテーション開始直後に評価，プレゼンテーションの終わりに近づいてから評価など，どのタイミングで評価する傾向があるかを表現できる．**図 1** (e-1)の評価者はプレゼンの途中で評価を完了し tM=4.44(分)，(e-2)の評価者はプレゼン時間の全体にわたって評価を行い tM=5.91(分)，(e-3)の評価者はプレゼン終了時にまとめて評価を入力し tM=10.2(分)，(e-4)の評価者はプレゼン時間中に評価を行わず（時間が経ってから入力し） tM=62.5(分)と，タイミングの違いが現れる．ただし，前述の評点の場合(sM)と同様，この平均(tM)からのみからでは，評価者に短時間に評価を行ったかどうかは表現できない点に注意が必要である．

評価時のタイムスタンプの標準偏差(Standard deviation of the evaluation time stamp: tSD) – 図 1(f)

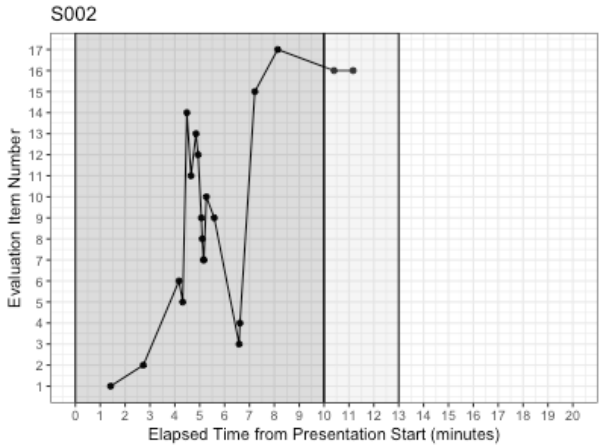
評価時のタイムスタンプの標準偏差(tSD)は，評価者が評価中に均等に時間を費やしたか，短時間で評価を行ったかを表現できる．**図 1** (f-1)の評価者はプレゼンテーション時間全体に渡って評価を行い tSD=3.74(分)であったのに対し，(f-2)の評価者は短時間に多数の項目の評価を入力したため tSD=0.24(分)と(f-1)より小さい値であった．

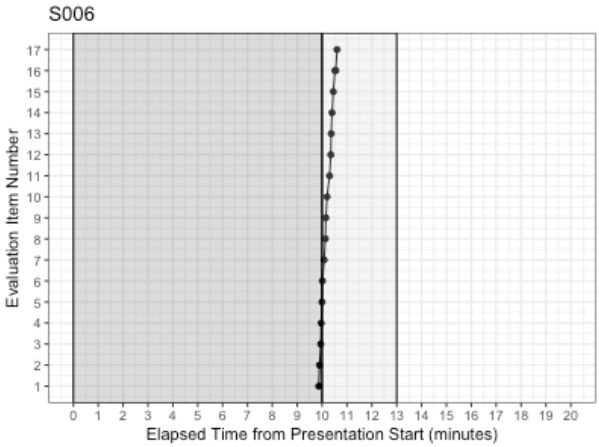




(a-1) 時間をかけて評価 (ET=12.3)　　**(a-2) 短時間で評価** (ET=0.73)

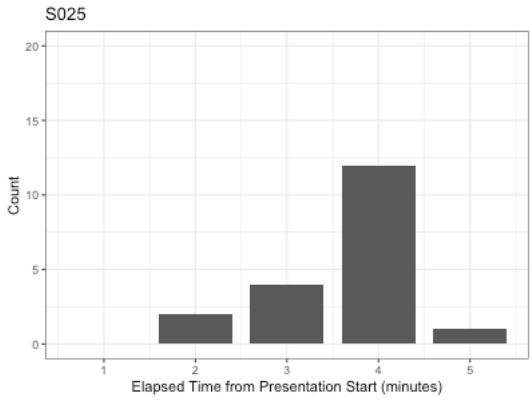
(a) 評価にかけた時間 (Evaluation Time: ET, 単位: 分)

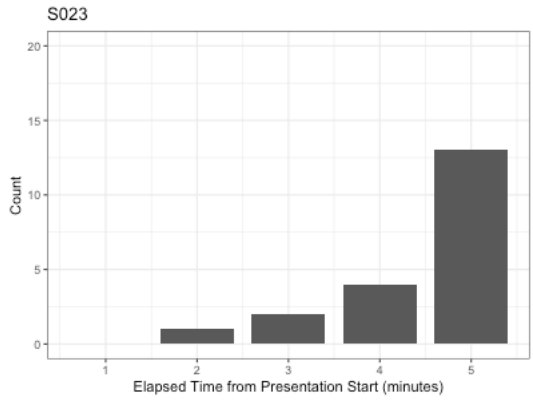




(b-1) 評価項目数以上の回数をクリック (CC=21)　　**(b-2) ほぼ最低限の回数のみクリック** (CC=18)

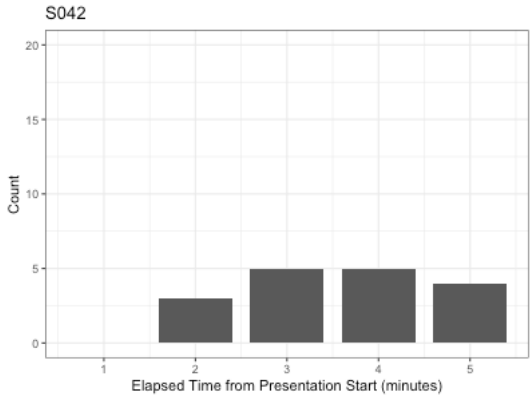
(b) クリック回数 (Click Count: CC, 単位: 回, 評価項目数: 17)



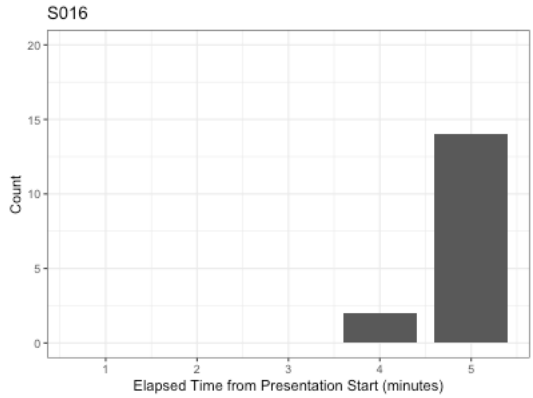


(c-1) 低い評点も使用(厳しい)(sM=3.63)　　**(c-2) 高い評点を多く使用(甘い)**(sM=4.45)

(c) 評点の平均 (Mean of the score: sM, 単位: 点, 満点: 5 点)

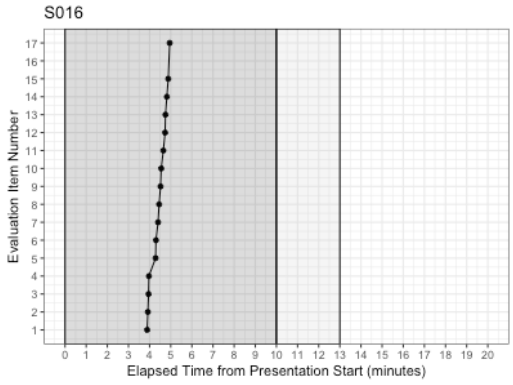


(d-1) 様々な評点を使用（sSD=0.99）

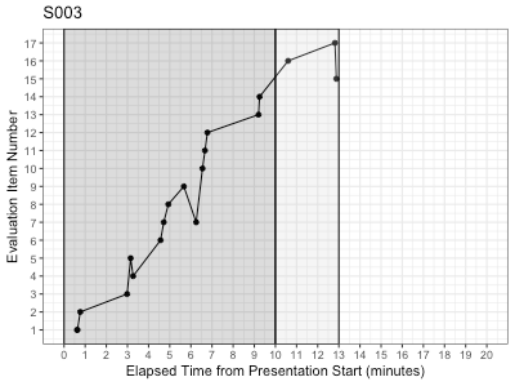


(d-2) 同じ評点ばかりを使用（sSD=0.34）

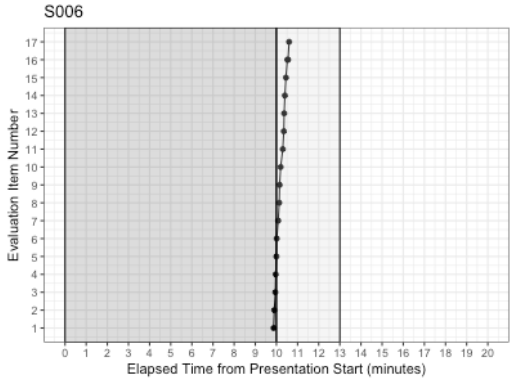
(d) 評点の標準偏差 (Standard deviation of the score: sSD，単位：点，満点: 5 点)



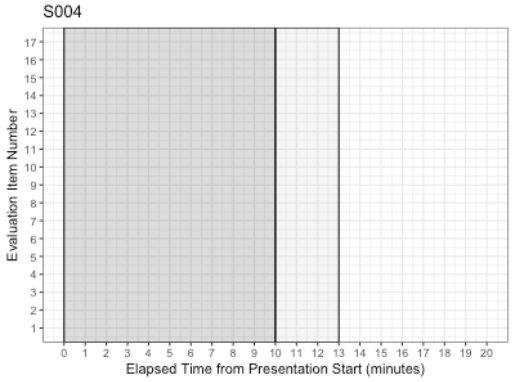
(e-1) プレゼンの途中で評価を完了（tM=4.44）



(e-2) プレゼン時間の全体にわたって評価（tM=5.91）



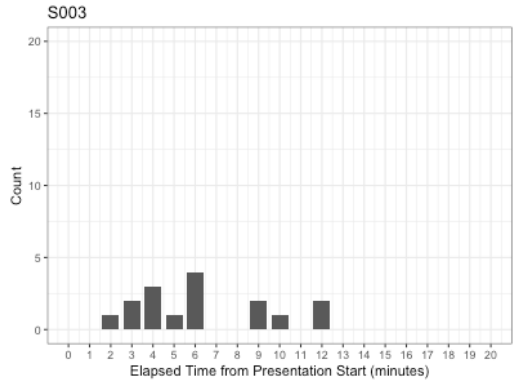
(e-3) プレゼン終了時にまとめて評価（tM=10.2）



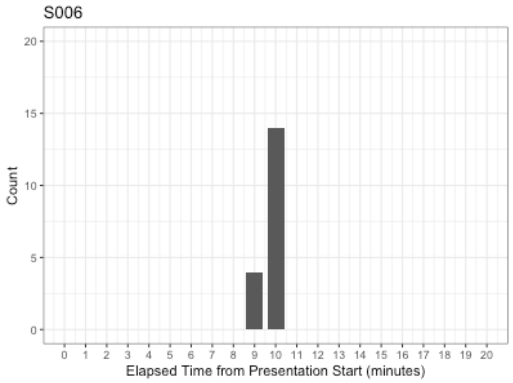
(e-4) プレゼン時間中に評価しなかった（tM=62.5）

(e) 評価時のタイムスタンプの平均* (Mean of the evaluation time stamp: tM，単位：分)

*評価項目クリック時のプレゼンテーション開始時からの平均経過時間



(f-1) プレゼン時間の全体にわたって評価（tSD=3.74）



(f-2) 短時間に多数の項目を評価（tSD=0.24）

(f) 評価時のタイムスタンプの標準偏差** (Standard deviation of the evaluation time stamp: tSD，単位：分)

**評価項目クリック時のプレゼンテーション開始時からの経過時間の標準偏差

図１ ６種の特徴量変数に関連する今回得られた特徴的な評価行動

3.3 RQ２：特徴量変数によって連続評価から受ける影響に差異があるか

3.3.1 各特徴量変数の連続評価の変化（予想）

第1章で述べたように，アンケートの回答行動分析の分野では「アンケート回答の繰り返しに伴い，回答者の正確に答える動機づけが低下し，短時間に回答する行動が増加する」という知見[11]がある．このため，相互評価の評価行動においても連続評価の実施により評価行動が変化し，それが特徴量変数に現れると考えられる．アンケートの回答行動の場合のように，連続評価により動機付けの低下等が生じると仮定したとき，予想される各特徴量変数の変化を整理したものが**表２**である．

表２ 連続評価時の各特徴量変数の変化（予想）

特徴量変数	連続評価時の変化	傾向
評価にかけた時間 (ET)	短時間で評価するようになる	減少
クリック回数 (CC)	最低限の回数しかクリックしないようになる	減少
評点の平均 (sM)	高い評点を使うようになる（甘くなる）	増加
評点の標準偏差 (sSD)	同じ評点ばかり使うようになる	減少
評価時のタイムスタンプの平均 (tM)	後からまとめて評価するようになる プレゼン開始前に終わらせてしまう	不明
評価時のタイムスタンプの標準偏差 (tSD)	短時間に多数のクリックをするようになる	減少

また，この変化を連続評価に伴って減少するか増加するかの観点で要約したのが**表２**の最右列である．評価にかけた時間（ET），クリック回数(CC)，評点の標準偏差(sSD)，評価時のタイムスタンプの標準偏差(tSD)の４変数は連続評価に伴って値が減少すると予想した．一方で，評点の平均(sM)は連続評価に伴って値が増加すると予想した．評価時のタイムスタンプの平均(tM)については，連続評価の影響を受けると考えられるが，予想される変化の方向が一定ではないため，明確な傾向は現れないと予想した．

3.3.2 各特徴量変数の連続評価の変化（実際）

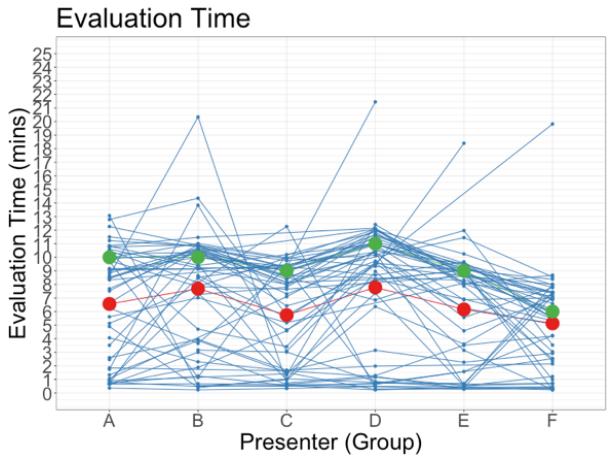
前項の予想を検証するため，**図２**に６グループに対する評価における各特徴量変数の推移を図示した．細線が各学習者の特徴量変数，赤色の大点は平均である．また，ET（**図２**（a-1), (a-2)）の緑点は当該グループのプレゼンテーションの長さ，CC（**図２**（b-1), (b-2)）の緑点は評価項目数を表す．ET と CC については，**図２**（a-1),(b-1)に実測値，**図２**（a-2),(b-2)に相対値を示している．ET の場合は各グループのプレゼンテーションの長さに対する ET の比，CC の場合は評価項目数に対する CC の比，つまり１項目あたり平均何回クリックしたかを示している．

グラフ形状からの議論

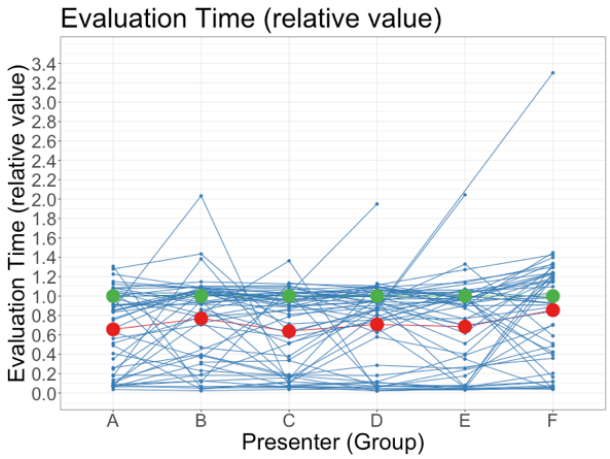
ET について，**図２**（a-1)では一見連続して評価を行うと ET が短くなっているように見える．しかし，プレゼンテーションの長さを表す点に着目すると，ET の傾向は，プレゼンテーションの長さの際の影響を大きく受けていることがわかる．この影響を除くために，各グループのプレゼンテーションの長さに対する ET の比を算出し，可視化した図 2 (a-2)では，**図２**（a-1)で見られた連続した評価に伴う下降傾向が消滅している．以上より，グラフ形状からは，「評価にかけた時間(ET)」はプレゼンテーションの長さの際の影響を大きく受け，また連続して評価を行ってもクラス全体として単純に短くなっていくことはないと言える．

図２（b-1),(b-2)から，CC においても連続評価によって単純に減少することはないことがわかる．また**図２**(b-2)を見ると，平均して１評価項目あたり 2～3 回クリックする学習者も見られる一方で，多くの学習者は各項目１回程度に集中しており， CC はそれほど学習者間でばらつきが大きくない．つまり，CC は繰り返しの評価や授業の終わりが近づくなどしてもあまり影響を受けず，評価者間でも差異が少ない可能性が示唆される．

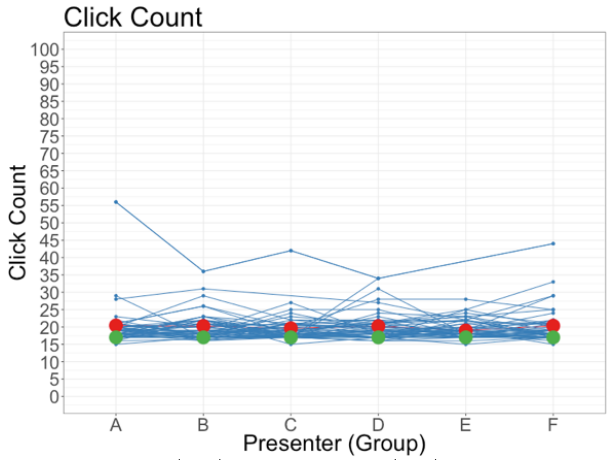
図２（c)より sM も先２種の変数と同様，連続評価を行なっても単純に増加することはないことがわかる．なお，評価対象によって sM の値の変動が見られるが，これは純粋に評価対象のプレゼンテーションの質を反映している可能性がある．しかしながら，今回は教員等の基準となる評価を行っていないため，検証できない．このような分析については，過去に別報告[12]において学習者と TA の評価の比較を行っている．



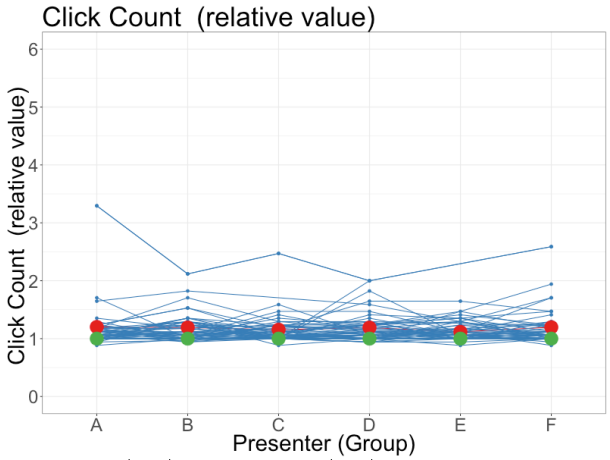
(a-1) 評価にかけた時間（ET）



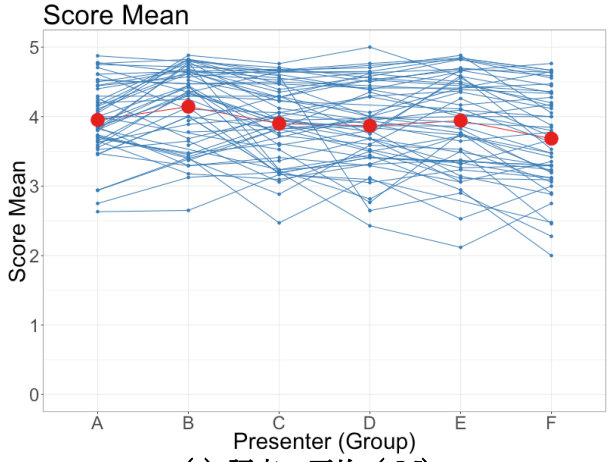
(a-2) 評価にかけた時間(ET) 相対値



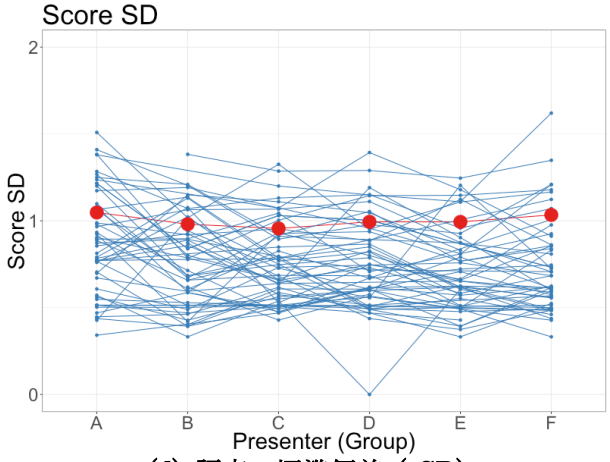
(b-1) クリック回数（CC）



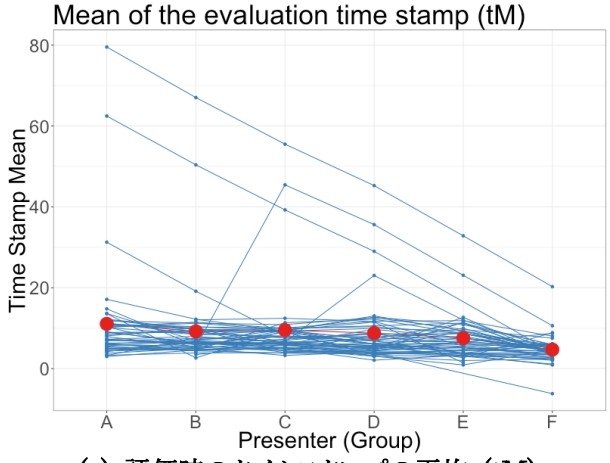
(b-2) クリック回数(CC) 相対値



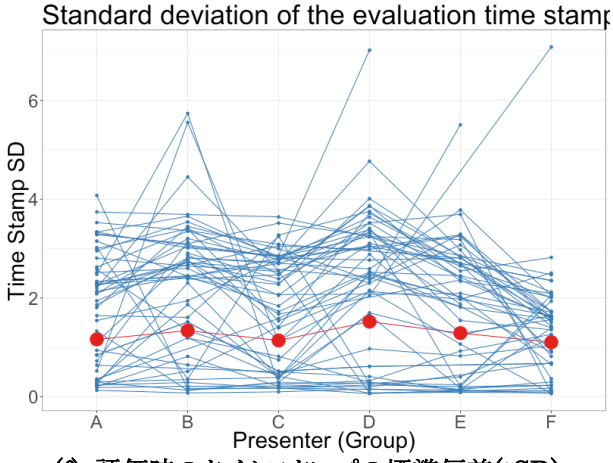
(c) 評点の平均（sM）



(d) 評点の標準偏差（sSD）



(e) 評価時のタイムスタンプの平均（tM）



(f) 評価時のタイムスタンプの標準偏差(tSD)

図 2 連続評価時の特徴変数の推移

残る sSD，tM，tSD についても連続評価に伴う明確な増加・減少は見られない．以上より，グラフの形状のレベルではどの特徴量変数も，連続評価に伴う単純な減少・増加は見られなかった．一方で，評価者間でのばらつきが大きい変数と小さい変数が見られた．

分散分析による議論

続いて，グラフの形状のレベルでは捕捉できない性質を議論するため，6種の特徴量変数に対し，連続評価の影響を分析するため反復測定の一元配置分散分析を行った．その結果，CC を除く ET, sM, sSD, tM, tSD において評価対象ごとの特徴量変数に有意な差が見られた（4変数とも $p<.001$）．一方で，CC については有意差はみられなかった（$p=.486$）．つまり，CC 以外の変数についてはグラフの形状のレベルでは傾向は見られないが，評価対象ごとに差異が見られることが明らかになった．また，CC については，グラフからも示唆されるように，繰り返しの評価や授業の終わりが近づくなどしてもあまり影響を受けず，評価者間でも差異が少ないことが明らかになった．

4. おわりに

本稿では，評価行動の特徴を定量的に議論するため，評価行動の特徴を定量的な特徴量変数として抽出し，その傾向や性質を明らかにすることを目的とした．

「RQ１：特徴的な評価行動は定量的に表現可能か」については，６種の特徴量変数により，関連する評価行動の特徴を捉え，表現することができた．また，「RQ２：特徴量変数によって連続評価から受ける影響に差異があるか」については，グラフの形状に現れるレベルではどの特徴量変数も，連続評価に伴う単純な減少・増加は見られなかった．一方で，評価者間でのばらつきが大きい変数と小さい変数が見られた．特にクリック回数については，分散分析においても評価者間で差異が少ないことが明らかになった．

今後課題として，今回のデータを縦断データと捉えて潜在成長モデルを用いて，個人内変化の傾向と変化の個人差について議論を行いたい．

参 考 文 献

(1) 藤原康宏, 大西仁, & 加藤浩: “学習者間の相互評価に関する研究の動向と課題”, メディア教育研究, Vol.4, No.1, pp.77-85 (2007)

(2) 深澤真: “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, Vol.21, pp.181-190 (2010).

(3) Orsmond, P., Merry, S., & Reiling, K.: “The importance of marking criteria in the use of peer assessment”, Assessment & Evaluation in Higher Education, Vol.21, No.3, p.239-250 (1996)

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

(5) Brown, J. D.: New Ways of Classroom Assessment. New Ways in TESOL Series II, Innovative Classroom Techniques. TESOL, (1998)

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

(7) Hughes, I. E., & Large, B. J.: “Staff and peer-group assessment of oral communication skills”, Studies in Higher Education, Vol.18, No.3, pp.379-385 (1993)

(8) Freeman, M.: “Peer assessment by groups of group work”, Assessment & Evaluation in Higher Education, Vol.20, No.3, pp.289-300 (1995)

(9) 藤原康宏，大西仁，加藤浩：“公平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」”，日本教育工学会論文誌，Vol.31，No.2，pp.125-134 (2007)

(10) Couper, M. P.: “Web surveys: A review of issues and approaches”, The Public Opinion Quarterly, Vol.64, No.4, pp.464-494 (2000a)

(11) Yan, T., and tourangeau, R.: “Fast times and easy questions: the effects of age, experience and question complexity on web survey response times”, Applied Cognitive Psychology, Vol.22, No.1, pp.51-68 (2008)

(12) Horikoshi, I., and Tamura, Y.: “Analysis of “Evaluation Behavior” Using Students’ Peer Assessment Process Data”, The 27th International Conference on Computer in Education (ICCE 2019), pp.755-758 (2019)