学生による相互パフォーマンス評価の信頼性
――多相ラッシュモデルを用いて――

石　川　勝　彦
百　瀬　光　一

はじめに

本研究は、学生同士によるパフォーマンス評価の信頼性を、多相ラッシュモデルを用いて評価することで、ピア評価の信頼性に影響を与える要因を検討することを目的とする。具体的には「評価者」の信頼性を改善することで、評価システム全体がどのように改善するか検討する。

学習者が同時に評価者となることの重要性が指摘されている。「学生の主体的な学び」の実装が教育における重要課題とされ、アクティブ・ラーニングへの転換が推進されている（文部科学省）。

アクティブ・ラーニングとして頻用される評価は、正答／誤答といった二分法ではなく、正答が一意に定まらないオープンエンドな評価である。アクティブ・ラーニングでは、ライティング、プレゼンテーション、課題解決、企画立案等を対象に、多様な評価の観点を運用して評価活動が行われる。具体的にはルーブリックを運用した評価活動が一般化しつつある。ルーブリックには到達目標と評価観点（および評価基準）が示されており、評価者はルーブリックを用いて評価を進める。

ルーブリックを用いた相互評価の教育的意義は、「評価者」の視点を取得することと深いかかわりがある。同一の課題に他者がどのように取り組んだのか観察し、自己の取り組みと比較するなかで、視点や方法の多様性や代替性に目を開いていくことを狙いの一つとしていると言えるだろう。評価能力が成熟することがパフォーマンスを向上させる性質があるならば、学習者を、併せて評価者としても熟達させることは重要な教育のチャンネルとなりうる。

ルーブリック評価が適合する場面は、正答が一意に定まる多肢選択式のテストが利用できない場面、つまりパフォーマンス評価が要求される場面である。

当然のことながらルーブリックには評価の観点が記載されている。このことは些細ではあるが非常に重要と思われる。熟達者が固有に持ちうる評価の視点を初学者が自力で獲得することには困難が伴うと思われる。ルーブリックに記載される評価の観点は、自力では獲得困難な観点を先取り的に取得する機会であるとともに、同時に評価の視点であり、授業の到達目標でもある。ルーブリックは学習者にとって多機能的に作用する重要な教材と考えられる。またルーブリックがなければ観点も評価基準もバラバラになり、フィードバックを受ける学習者に「評価に客観性がない」などの不公正感を感じさせる危険も生じうる（もちろん評価基準が見えない中で評価を受けることの学習効果を否定するものではない）。

その一方、ルーブリックを用いたからといって、直ちに、評価に「客観性」を実装できるわけではない。一般に測定を行う際には、信頼性（reliability）および妥当性（validity）のチェックが必要となる。例えば古典的テスト理論における信頼性概念は、「測定誤差が小さいこと」と要約できる（岡田、2011）。例えば平行テスト法（真値が等しく、また誤差の分散も等しいテストを２セット作成し、それぞれの測定値を用いて信頼性を

求める方法）を用いた際に、２つのテストの相関係数は信頼性の指標となる（一方で平行テストを開発すること自体の困難が指摘されている）。妥当性とは「測定すべきものを測定している程度」と理解することができる。例えば基準関連妥当性（関連するテストと妥当な関連性がある）や構成概念妥当性（測定尺度の因子構造が概念レベルの議論と符合する）の観点から議論されてきた（村山、2012）。

ルーブリックによるパフォーマンス評価は「測定」の一種と理解することができるため、あらゆる測定と同様に、信頼性・妥当性が担保されていることが必要である。具体的にはルーブリックを構成する評価項目、および各評価項目を構成するカテゴリーの精度を分析し、問題が発見された場合には改善案を案出することにつなげることが望ましい。同時に、複数の評価者が存在する場合には、評価特性が適切（多くの評価者が卓越した受験者を高く、平均的な受験者をそれよりも低く評価するなど）であることが望ましい。

本研究は、学生による相互パフォーマンス評価の信頼性を、多相ラッシュモデルを用いて検討することで、ピア評価の信頼性に影響を与える要因を検討することを目的とする。具体的には「評価者の信頼性」を改善することで、評価全体がどのように改善するか検討する。

本稿では、具体的には大学の教職科目である「教職実践演習」におけるプレゼンテーション・ソフトを用いた模擬授業に対し、学生に相互パフォーマンス評価を行わせ、３相データ（受験者、項目、評価者）を取得した。具体的には、受験者はプレゼンテーション・ソフトを用いた模擬授業を行った受講生、項目はプレゼンテーションを用いた模擬授業の質を評価する４項目、評価者は模擬授業を評価した受講生であった。

教職実践演習は、４年次の教職実習後、教育実習の振り返りおよび教職課程全体のまとめ科目に位置付けられている。その狙いは、教育実習及び

その後の指導を通して明らかになった課題を重点的に確認・指導することである。

相互パフォーマンス評価データを教職実践演習で取得した狙いは、教職課程を「修了」する時点における評価者としての受講者たちの評価の傾向を確認するとともに、模擬授業の評価項目の項目反応特性を確認することにある。分析の狙いであるが、第一に、もし評価者である受講生たちが、ラッシュモデル（後述する）に適合した評価者特性を示すならば、4年間の教職課程を通じて一定程度の評価能力を獲得していると考えることができる。第二に評価者たちがラッシュモデルに適合した評価者特性を示さないならば、評価特性の不良な評価者と良好な評価者を分離して再度モデリングし、評価者特性が評価全体に与える影響を確認する。第三に、評価項目の項目反応特性を確認し、適切な困難度、識別力、適合性を示すかどうか検討する。もし項目反応特性に不良な点がみられた場合は、原因を究明するとともに、項目の再整備に向けて考察を行う。

分析の具体的な手続きを示す。受講生たちの評価者としての特性に注目する。具体的には、評価者特性がラッシュモデルにどの程度フィットしているか、フィットを改善した場合に受験者特性や項目特性の推定がどのように改善されるか確認する。具体的には、まず受講生の評価者特性を確認する。次に、ラッシュモデルにフィットしない評価者を削除して再分析を行い、受験者特性や項目特性がどのように変化するか確認する。このような手続きを通して、受験者特性とフィット、評価項目の特性とフィットがどのように改善されるのか観察する。

方法

調査方法・調査協力者

A 大学 4 年生を対象とする「教職実践演習」における「プレゼンテーション・ソフトを用いた模擬授業開発」の時間を利用した。当該時間では、13名の学生が模擬授業を行い、これを21名の学生が評価した。21名のうち13名は模擬授業の実演者であった。模擬授業はおよそ 5 分の持ち時間の枠内で行われた。

評価項目

模擬授業の評価は以下の 4 つの項目を用いて行われた。「 1 時間の授業の到達目標が明確に示されていた（purpose clarity）」「発問・指示・説明が明確だった（question clarity）」「図・写真などが、適切な場面で、適切に提示されており、授業資料の構成が効果的だった（document appropriation）」「何のために（目標）、何を（内容）、どのように（方法）作業させるか、教材作成が適切だった（material appropriation）」。いずれも「 5 . 当てはまる～ 1 . 当てはまらない」の 5 件法で評価された。

分析

多相ラッシュモデル（e.g. Eckes, 2015）を用いて分析した。ソフトウェアは Minifac（Linacre, 2019）を用いた。多相ラッシュモデルは受験者能力、項目難易度に加えて任意に相（facet）を追加し、ロジッツ（logits）を単位として統一的に要素の難易度や能力を評価することができる。本研究では、評価者（rater）が21名存在することから、受験者能力、項目難

易度に加えて、評価者を加えた3相ラッシュモデルを推定することとした。評価者の相はロジッツ上で各評価者の厳しさ（severity）、甘さ（leniency）を評価するものである。

多相ラッシュモデルでは、各相の推定結果を、ロジッツとよばれる推定値に一元的に表現することができる。ロジッツの値が高くなるほど受験者能力は高く、評価者は厳しく、項目難易度は高いことを表す。逆にロジッツの値が小さいほど、受験者能力は低く、評価者は甘く、項目難易度は低いことを表す。

加えて、各相について「適合度」を評価し、測定の妥当性をチェックする。適合度とは、ラッシュモデルが仮定しているガットマン状態にデータが適合している程度のことである。ガットマン状態とは、「項目」を例にとった場合、難易度の低い項目ほど正答率が高く、難易度の高い項目ほど正答率が低い状態としてイメージできる。さらに、テスト全体でみれば、受験者能力の低い受験者ほど正答率が低い項目が多く、受験者能力の高い受験者ほど正答率の高い項目が多い状態である。加えて、受験者能力の低い受験者が正答するが、受験者能力の高い受験者が正答しない項目が「存在しない」ことが重要な条件である。このような一連の仮定を満たしたスコアのマトリックスがガットマン状態であり、適合度とは実際に得られたマトリックスがガットマン状態に近似している程度を評価する指標である。より具体的には、各相の実測値が対応する相の期待値と適合する程度を表す。実測値とは評価者が各項目について受験者に与えたスコアである。期待値とは、同一の受験者に対し項目の困難度、受験者のパフォーマンスが与えられている時に、評価者が与えるであろう評価である。実測値と期待値の間に大きなギャップがある場合、評価の信頼性に問題が存在する可能性がある。

乖離信頼性統計量（Separation reliability statistic）はクロンバックの α

係数と同様 0 ～ 1 の値をとる。それゆえ評価者の Separation が低いことは望ましい。なぜなら評価者間で評価にバラつきがあることは望ましくないからである。Separation は他の相、すなわち受験者能力、項目困難度についても計算される。

加えて項目の評価尺度の適合性を評価するために、評価尺度カテゴリー機能（the functioning of the rating scale categories）を調べた。評価尺度カテゴリー（本研究では 5 件法）はテストの信頼性に影響する可能性がある。それゆえ評価者はすべてのカテゴリーに平等に注意を払うことが必要かつ重要である。少なくとも、どのカテゴリーも10回以上選択されることが必要である。カテゴリー間で生起数に大きな差がある場合、カテゴリーが不適切に作用している可能性がある。それゆえ、各カテゴリーが均一に使用されていることは望ましい。加えて、各カテゴリーの使用頻度の平均値は、アウトフィット値が 2 以下となり、使用頻度が単純直線的に増加することが望ましい。

本研究では、分析 1 では、13名の受験者、21名の評価者、 4 つの項目の 3 相ラッシュモデリングを行い、各相の各要素についてロジッツを算出するとともに、適合度を確認し、測定の信頼性を観察する。

分析 2 では得られた統計量に基づき、適合度の低い評価者を分析から削除し、受験者の能力推定と適合度、項目の難易度を適合度がどのように変化・改善されるかを確認する。

分析 1

1085の反応を分析の対象とした。

結果と考察

変数マップ

Figure 1 に変数マップを示した。変数マップは受験者特性、評価者の厳しさ、項目の困難度が同一のロジッツ尺度（一番左側のラカム「Measr」）の上に表現されている。それゆえ、すべての相は共通の尺度上で互いに比較することができる。次のカラム（「persons」の列）において、それぞれの数字は受験者の特性を表している。このカラムでは、それぞれの数字は受験者を表す。「平均的な」能力の受験者はロジッツ尺度の上で 0 に位置付けられるので、プラスの値の方向に数字が大きいほどその受験者の能力は高く、マイナスの方向に位置付けられるほどその受験者の能力は低いことを表す。 3 番目のカラム（「Rater」）は評価者の相を表す。厳しい評価基準を持っている評価者はカラムの一番上に位置づき、評価が甘い評価者はカラムの一番下に位置づく。項目の相は 4 番目のカラムに表現されている。各項目の平均的な困難度がロジット尺度上に表現されている。一番難しい項目はラカムの一番上に、一番容易な項目はカラムの最も下に位置づいている。 5 番目のカラムには項目の相に含まれる 4 項目を構成要素とする尺度平均が表現されている。

受験者の相

変数マップの左から 2 番目のカラムには受験者特性が示されている。受験者12が最も高く評価され、受験者10、受験者 3 、と続いている。一つ目の特徴として、受験者10、受験者 3 から受験者 4 までは切れ目のない連続的な評価が実現しているが、受験者12が受験者10、受験者 3 から離れている点がある。これらの受験者の間には他の受験者が位置付けられておらず、従って、切れ目のない連続的な評価が実現していないことが見て取れる。

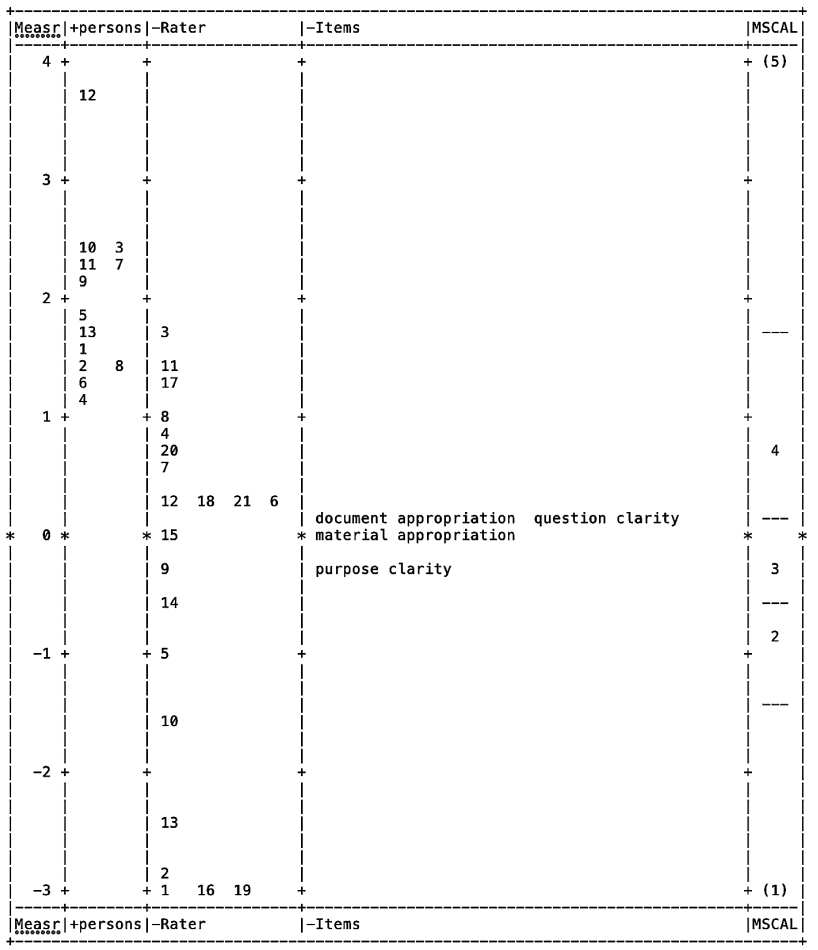

Figure1　変数マップ　変数マップは多相ラッシュモデルから生成される。3相（persons、Rater、Items）から成り、5件法で4項目、評価者は21名、評価対象は13名のモデルが推定された

第２に、最も評価の低い受験者４はロジッツ＝1.21である（Table １）。平均的な受験者がロジッツ＝0となっていないことから、受験者集団は全体的に高く評価されていることが伺える。

Table1　受験者特性

			Infit		Outfit	
person	Measure	SE	MnSq	Zstd	MnSq	Zstd
1	1.56	0.16	0.70	−1.30	0.88	−0.40
2	1.41	0.15	0.76	−1.00	0.77	−0.90
3	2.43	0.22	1.04	0.20	1.57	1.60
4	1.21	0.14	0.84	−0.70	1.07	0.30
5	1.89	0.19	0.87	−0.40	0.77	−0.80
6	1.25	0.15	1.05	0.20	0.92	−0.20
7	2.29	0.21	0.78	−0.80	0.75	−0.70
8	1.46	0.16	0.98	0.00	0.94	−0.10
9	2.08	0.20	1.50	1.70	2.03	2.80
10	2.40	0.23	0.79	−0.70	1.05	0.20
11	2.36	0.22	1.46	1.60	1.38	1.10
12	3.77	0.34	2.40	3.60	2.15	1.60
13	1.70	0.17	0.48	−2.50	0.44	−2.60
Mean	1.99	0.20	1.05	0.00	1.13	0.10
SD	0.67	0.05	0.48	1.50	0.49	1.40
Separation	3.12					
Strata	4.49					
Reliablity	0.91					
χ^2	129.80 ***					

$^{**}p < .01$, $^{*}p < .05$, $^{+}p < .10$

評価者の相

変数マップには評価者の相対的な厳しさが表現されている。評価者３が最も厳しい評価者であり、ロジッツ＝1.72だった。評価者１、16、19が最も評価が甘く、ロジッツはそれぞれ-4.80だった（Table ２）。変数マップに示されている通り、評価者の間には6.51ロジッツの厳しさの差異が存在する。この散らばりは統計的に有意であった（χ^2=285.50、p＜.001）。評価者の厳しさには有意な違いが存在すると言える。Table ５に示されているようにSeparation=2.52、Separation reliability=0.86となっており、各評価者の評価の仕方は類似しておらず、その厳しさにはバラつきがあることが示されている。変数マップからもこの知見はサポートでき、具体的には評価者は同じロジッツのレベルには位置付けられていないことが伺える。

インフィットおよびアウトフィット統計量はどの評価者がラッシュモデルに適合しているかを表している。インフィット統計量はMnSq（Mean Squared）が0.17～2.47に分布しており、平均は1.04、SDは0.45だった。アウトフィット統計量はMnSqが0.2～2.48に分布し、平均は1.14、SDは0.58だった。フィット統計量は0.5～1.5の時に、評価がモデルの予測値に適合していることを表す。平均値はやや高めと言える。インフィットもしくはアウトフィット統計量が0.5～1.5の外にある評価者は21名中８名だった。このことが表すのは、これらの評価者は難しい項目に対し高いスコアリングを行っている、あるいは受験者能力の低い受験者に高いスコアを与えている可能性を示している。同時に、容易な項目に低いスコアを与えている、あるいは受験者能力の高い受験者に低いスコアを与えている可能性がある。

受験者と評価者の全体的な配置を見てみると、受験者は１以上のロジッ

Table2　評価者特性

			Infit		Outfit	
rater	Measure	SE	MnSq	Zstd	MnSq	Zstd
1	−4.80	1.83				
2	−2.84	0.72	0.98	0.10	1.05	0.30
3	1.72	0.13	1.05	0.30	1.03	0.20
4	0.92	0.18	0.68	−1.10	0.84	−0.50
5	−0.99	0.33	2.47	3.90	2.45	3.20
6	0.22	0.23	1.43	1.40	1.65	2.20
7	0.64	0.20	0.76	−0.80	0.73	−1.00
8	1.03	0.17	0.17	−4.80	0.20	−4.70
9	−0.36	0.27	1.17	0.70	0.98	0.00
10	−1.51	0.40	0.90	−0.20	0.78	−0.30
11	1.36	0.15	0.97	0.00	0.90	−0.30
12	0.32	0.23	1.23	0.80	1.03	0.10
13	−2.45	0.59	0.97	0.10	0.93	0.10
14	−0.62	0.30	1.12	0.50	1.87	2.40
15	0.05	0.24	0.77	−0.80	0.78	−0.80
16	−4.80	1.83				
17	1.24	0.16	0.58	−1.90	0.60	−1.80
18	0.27	0.23	1.10	0.40	1.07	0.30
19	−4.80	1.83				
20	0.72	0.20	1.28	0.90	2.48	4.10
21	0.27	0.23	1.08	0.30	1.06	0.30
Mean	−0.69	0.50	1.04		1.14	
SD	2.03	0.56	0.45		0.58	
Separation	2.52					
Strata	3.69					
Reliablity	0.86					
χ^2	285.50 ***					

$^{**}p < .01$,　$^{*}p < .05$,　$^{+}p < .10$

ツに偏在しており、一方評価者は2ロジッツ以下に偏在しているため、受験者の相と項目の相に「重なり」がほとんどみられない。つまり、多くの評価者は、評価が甘すぎるがゆえに受験者能力を適切に識別できないと考えることもできる。

項目の相

項目の相の推定結果も同じく変数マップで確認できる。項目難易度は-0.25～0.11ロジッツに分布している。最も困難度が高かったのは「発問の明確さ」で、もっとも容易だったのは「到達目標の明確さ」だった。

$\chi^2=7.60$（$p<.05$）と有意な項目難易度のバラつきが示唆されている（Table 3）。しかしながら、Separation=1.06、Reliability=0.53であったので、項目難易度のばらつきはそれほど大きくないと考えることができる。

Table3　項目特性

			Infit		Outfit	
Item	Measure	SE	MnSq	Zstd	MnSq	Zstd
purpose clarity	-0.25	0.11	1.10	0.70	0.92	-0.30
question clarity	0.11	0.10	0.79	-1.60	1.24	1.30
document appropriation	0.10	0.10	1.09	0.70	1.29	1.60
material appropriation	0.04	0.10	0.82	-1.30	1.09	0.50
Mean	0.00	0.10	0.95	-0.40	1.13	0.80
SD	0.15	0.01	0.15	1.10	0.15	0.80
Separation	1.06					
Strata	1.75					
Reliablity	0.53					
χ^2	7.60	*				

$^{**}p<.01$, $^{*}p<.05$, $^{+}p<.10$

Stara＝1.75であるので項目困難度はおおよそ１～２層に収まっている。逆に言えば項目の難易度の類似性が高く、多様性に欠けると言える。フィット統計量をみてみると、4項目とも0.5～1.5に収まっており、ラッシュモデルの予測値と実測値のズレは問題にならない範囲に収まっていると考えられる。

いずれの項目もおおよそロジット０の周辺にばらついている。したがって、高い能力を示す受験者を平均的な受験者から識別する性能を備えていないと考えることもできる。

評価尺度カテゴリー機能

５件法の選択肢がどれくらい良好に評価者に理解され、識別力のある評価に資するものであったのかを調べるため、評価カテゴリーの機能を分析した（Table ４）。目的は最適な評価カテゴリー数を実証的に決めることである。この結果によると各カテゴリーが選択された頻度は単純増加していないことが見て取れる。Average measure がマイナス（-0.05ロジッツ）を示しているカテゴリー２（weak）が存在していることがこれを示唆している。さらにアウトフィット値が0.5～1.50をはみ出しているカテゴリーがある（カテゴリー1、very weak およびカテゴリー２、weak）。

Table4　評価尺度カテゴリー機能

Category	Total	Percentage	Average measure	Expected means	Outfit MnSq	Response Category Name
1	14	2%	0.44	0.06	3.00	very weak
2	16	2%	-0.05	0.33	0.40	weak
3	56	6%	0.73	0.72	1.00	average
4	332	36%	1.31	1.33	1.00	good
5	667	55%	2.65	2.64	1.00	very good

このことは測定のプロセスにノイズが混じっていることを意味している。カテゴリー数は4件法もしくは3件法に縮約したほうが適切となる可能性がある。

以上の統計量から、甘い評価者が多い傾向、項目難易度が類似しており識別力に問題が残る点が指摘される。

分析2では、問題のある評価者がフィット統計量からあぶりだされていることに基づき、これらの評価者による評価データをデータセットから削除しその上で分析1と同様に3相ラッシュモデルの推定結果を解釈する。評価者の削除により、受験者能力、項目難易度、評価カテゴリーの推定値やフィットがどのように改善されるか検討することとする。

分析2

分析2では、分析1においてフィット統計量から評価スコアがラッシュモデルに対して適合が良くないと判断された評価者を削除し、適合が優れている，または許容範囲内にある評価データのみを用いて3相ラッシュモデルを推定し、受験者特性、項目難易度、評価カテゴリーの推定値がどのように変化するか、フィットが改善されるかを確認することを目的とする。

結果と考察

評価者のインフィット値、アウトフィット値が0.5～1.50に収まらない評価者をデータから削除し、3相ラッシュモデルを推定し、さらに評価者のインフィット値、アウトフィット値を確認する、という分析プロセスを繰り返したところ、9名の評価者が削除対象となり、12名の評価者による評価データが分析対象となった。データセット全体のインフィット値は

0.71～1.10、アウトフィット値は0.74～1.07の範囲に収まった。621の反応が分析対象となった。

変数マップ

Figure 2 に変数マップを示す。分析 1 と同様に、ロジッツ、受験者、評価者、項目、尺度の順にカラムが構成されている。以下、各相に分けて、

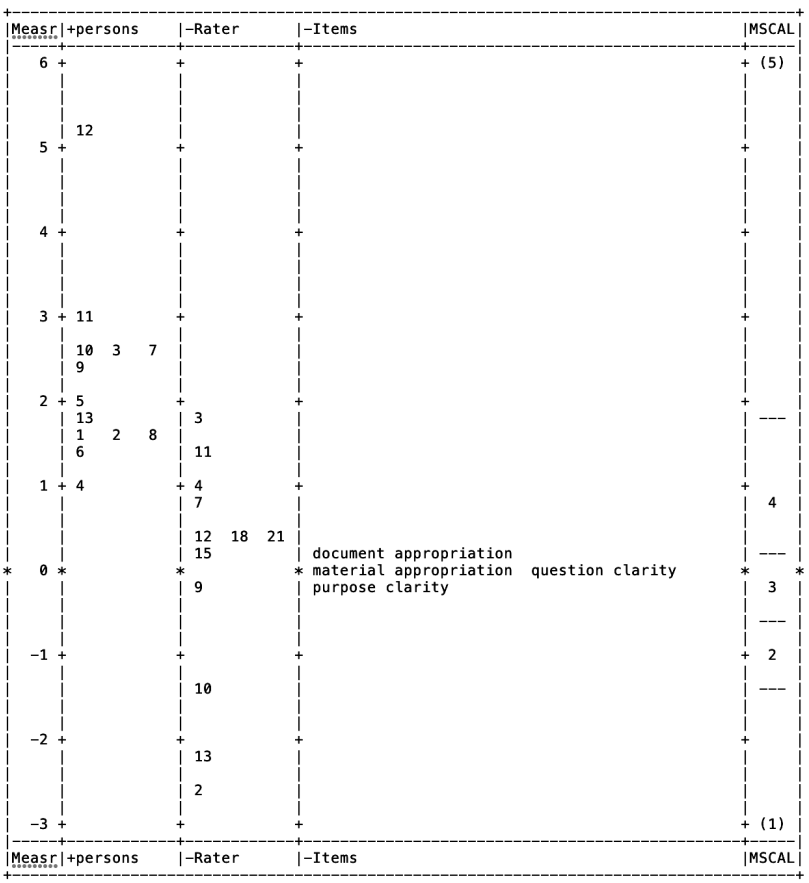

Figure2　variable map

評価者の適合度を改善したことでどのような変動や改善がみられたか確認する。

受験者の相

ロジッツのレンジは、分析１では1.21～3.77で2.56、分析２では

Table5　受験者特性

			Infit		Outfit	
person	Measure	SE	MnSq	Zstd	MnSq	Zstd
1	1.66	0.20	0.69	−1.20	0.91	−0.10
2	1.62	0.20	0.93	−0,10	0.71	−0,90
3	2.56	0.27	1.22	0.70	2.03	1.90
4	1.01	0.17	0.76	−1.00	0.90	−0.20
5	2.01	0.23	1.15	0.50	0.94	0.00
6	1.33	0.18	1.19	0.80	0.96	0.00
7	2.64	0.28	0.87	−0.30	0.68	−0.60
8	1.5	0.19	1.10	0.40	1.03	0.20
9	2.42	0.25	1.24	0.80	0.96	0.00
10	2.56	0.27	0.82	−0.40	1.15	0.40
11	2.95	0.31	1.13	0.40	0.87	0.00
12	5.14	0.72	0.87	0.00	0.34	0.00
13	1.78	0.21	0.48	−2.20	0.40	−2.20
Mean	2.24	0.27	0.96	−0.10	0.91	−0.10
SD	1.01	0.14	0.23	0.90	0.39	0.90
Separation	3.18					
Strata	4.57					
Reliablity	0.91					
χ^2	98.90 ***					

$^{**}p < .01$, $^{*}p < .05$, $^{+}p < .10$

1.5～5.14で3.64となった（Table 5）。受験者能力のバラつきを表す指標をみてみると、分析1では Separation=3.12、Strata=4.49、Reliability=0.91であり、分析2では Separation=3.18、Strata=4.57、Reliability=0.91であった。バラつきの程度に大きな変化はみられないと考えてよいだろう。ロジッツの平均値は、分析1では1.99、分析2では2.24となった。フィット統計量をみてみると、分析1において不適合な受験者（インフィットないしアウトフィットの MnSq が0.5～1.5に収まっていない）は4名だったのに対し、分析2では3名だった。結果的に、受験者能力は分析1から分析2にかけてより「高く」推定され、受験者のラッシュモデルへの適合度に大きな改善は見られない結果となった。受験者能力のバラつきにも変化はみられなかった。

評価者の相

ロジッツのレンジは、分析1では-4.80～1.71で6.51、分析2では-2.67～1.83で4.50となった。Separation、Strata、Reliability は分析1ではそれぞれ2.52、3.69、0.86、分析2では3.78、5.35、0.93であった（Table 6）。分析2において分析1よりも評価者の厳しさのバラつきは大きくなったと考えることができる。ロジッツの平均値は、分析1では-0.69、分析2では0.00だった。インフィット値ないしアウトフィット値が0.5～1.5から外れた評価者を削除したため、こうした評価者はいない。フィットの不良な評価者を削除したことで、評価の厳しさのバラつきは小さくなり、全体として評価は厳しくなった。

項目の相

ロジッツのレンジは分析1では-0.25～0.11で0.36、分析2では-0.16～0.11で0.27だった。項目難易度のバラつきを表す指標をみてみる

Table6　評価者特性

			Infit		Outfit	
rater	Measure	SE	MnSq	Zstd	MnSq	Zstd
2	−2.67	0.72	0.96	0.10	1.03	0.40
3	1.83	0.14	1.02	0.10	1.05	0.20
4	1.01	0.18	0.71	−1.10	0.93	−0.10
7	0.75	0.2	0.78	−0.70	0.74	−0.90
9	−0.2	0.27	1.10	0.40	0.90	−0.10
10	−1.33	0.4	0.83	−0.40	0.69	−0.30
11	1.46	0.15	0.98	0.00	0.98	0.00
12	0.45	0.22	1.03	0.20	0,82	−0.50
13	−2.27	0.59	0.97	0.10	0.93	0.20
15	0.19	0.24	0.74	−0.90	0.79	−0.50
18	0.4	0.22	1.08	0,30	1,07	0.30
21	0.4	0.22	1.05	0.20	1.02	0.10
Mean	0	0.3	0.94	−0.10	0.91	−0.10
SD	1.34	0.17	0.13	0,50	0,12	0,40
Separation	3.78					
Strata	5.37					
Reliablity	0.93					
χ^2	183.10 ***					

$^{**}p < .01$,　$^{*}p < .05$,　$^{+}p < .10$

と、Separation、Strata、Reliability はそれぞれ分析 1 では1.06、1.75、0.53、分析 2 では0.00、0.33、0.00だった（Table 7）。項目難易度のバラつきは分析 2 よりも分析1の方で大きかったとみえる。ロジッツの平均値は分析 1 で0.00、分析 2 でも0.00だった。フィット値が不適切な値（0.5〜1.5）を示した項目は分析 1 にも分析 2 にも見られなかった。分析 2 において項目難易度のバラつきは縮小され、多様性はさらに失われた傾向がみられた。全体の難易度は同程度、フィットはともに良好であった。

評価尺度カテゴリー機能

分析 2 でも Average measure は単純増加の傾向を示さなかった（Table 8）。具体的にはカテゴリー 1 とカテゴリー 2 の Average measure が同程度となり、その後は単純増加となった。アウトフィット値が0.5〜1.5をはみ出したカテゴリーは消失し、フィットは改善した。

Table7　項目特性

			Infit		Outfit	
Item	Measure	SE	MnSq	Zstd	MnSq	Zstd
purpose clarity	−0,16	0.13	1.07	0.40	0.73	−0.70
question clarity	0.06	0.12	0.77	−1.40	0.95	0.00
document appropriation	0.11	0.12	1.20	1.20	1.28	0.90
material appropriation	−0.01	0.13	0.72	−1.80	0.69	−0.90
Mean	0,00	0.13	0.94	−0,40	0.91	−0.20
SD	0.10	0.00	0,20	1.30	0.23	0.80
Separation	0.00					
Strata	0.33					
Reliablity	0.00					
χ^2	2,70					

$^{**}p < .01$, $^{*}p < .05$, $^{+}p < .10$

Table8　評価尺度カテゴリー機能

Category	Total	Percentage	Average measure	Expected means	Outfit MnSq	Response Category Name
1	11	2%	−0.28	−0.13	0.70	very weak
2	12	2%	−0.29	0.21	0.20	weak
3	43	7%	0.83	0.68	1.20	average
4	190	31%	1.41	1.36	0.90	good
5	365	59%	2.99	3.02	1.10	very good

総合考察

フルデータによる３相ラッシュモデリング（分析１）、フィットの不良な評価者を削除したデータによる３相ラッシュモデリング（分析２）を比較し、受験者特性、項目特性、評価尺度カテゴリー機能に改善がみられるかどうか検討した。

評価者のフィットを改善した結果、受験者能力はより「高く」推定され、能力のバラつきとラッシュモデルへのフィットは変化しなかった。評価者の厳しさは厳しくなり、評価の厳しさのバラつきは小さくなった。項目の難易度は変化せず、難易度のバラつきは縮小され、フィットは変化しなかった。評価尺度カテゴリー機能は、いずれも単純増加しないカテゴリーを生じており、フィットは改善した。

注目すべき点として、第一に、受験者のフィットにほとんど変化がみられなかった。この結果から、もともと評価者たちは全体として良好な評価能力を有していたと考えることができる。８名のミスフィットを示した評価者たちの影響はもともと小さく、フィットを示した評価者たちの影響が頑健であったとみなすことが可能である。教職実践演習の受講生はすべて

教職実習を終えた4年生であったわけだが、この段階で、受講者はおおむね模擬授業をルーブリックを用いて評価するにあたり一定程度適切な評価スキルを獲得していたとみなすことができる。

第二に、項目難易度のバラつきが小さくなった。もともと項目難易度にばらつきが小さく、多様な難易度を有する項目セットを実現していない点が問題であったが、評価者のフィットを改善することで、項目間の難易度のバラつきがさらに小さくなった。本研究で用いた4項目は模擬授業を評価するにあたり難易度の同質性が高く難易度の多様性が不足している点に改善の余地を抱えていると思われる。

第三に、評価尺度カテゴリー機能において、フィットに一定の改善がみられた。具体的には、最も低いスコアを割り当てられているカテゴリーのフィットが改善した。評価者の評価スキルを改善することで、評価カテゴリーの機能性がラッシュモデルに適合しやすくなった。

変数マップから明らかになった、複数の相の間の関係性について整理する。第一に受験者能力に対し評価が甘い、第二にこれに付随して受験者能力と評価者の厳しさに対応がみられないロジッツを多く生じており識別力に問題を抱えている、第三に項目難易度が0ロジットに一元化されており各レベルの受験者を識別する項目セットを実現していない、と言える。

以上を踏まえ、今後の課題として、まず、項目数を増やすとともに項目の難易度に多様性を持たせることが必要である。併せてこれを通して項目の識別力の改善を目指す。次に、評価者の評価の厳しさを全体としてより厳しい水準に引き上げるとともに、厳しさのバラつきを小さくすることが必要である。

学習者のパフォーマンスは、学習者の自己評価能力と連動することが知られている（Kruger & Dunning, 1999）。自己評価能力が低い場合、当該人物のパフォーマンスは平均的な水準にとどまる傾向にあり、逆に評価能

力が高い場合には当人のパフォーマンスが高い水準にある傾向にある。学習者のパフォーマンスを向上するにあたり当人の自己評価能力を育成するアプローチを併せて与えることが重要となる可能性がある。

本研究では教職科目をリファレンスに、カリキュラム終了時点での評価者特性を確認するとともに、評価者のフィットを改善することで受験者特性、項目特性、評価尺度カテゴリー機能がどの程度改善するか、その影響力を調べることを目的とした。一定程度、適切な評価能力が獲得されていることがみてとれたが、評価者特性の改善により特に評価尺度カテゴリー機能が改善したため、評価能力のトレーニングも一定程度の教育効果を生ずる可能性が伺えた。今後は、より直接に、Kruger & Dunning（1999）の仮説が教育の場において妥当するか検証することも有用と思われる。自己のパフォーマンスに対する自己評価と他者評価を突き合わせ、濃密な振り返りを行うなどの学習を設けることで、パフォーマンス評価およびパフォーマンスがどのように改善するか検討する必要がある。

【引用文献】

Eckes, T. 2015 Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments. *Peter Lang Pub Inc.*

Kruger, J. & Dunning, D. 1999 Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77, 1121-1134.

Linacre, J. M. 2019 MINIFAC-Evaluation. Student and Demonstration（Demo）Version of FACETS. https://www.winsteps.com/minifac.htm

村山航　妥当性概念の歴史的変遷と心理測定学的観点からの考察．教育心理学年報，51，118-130.

文部科学省　教科等の本質的な学びを踏まえたアクティブ・ラーニングの視点からの学習・指導方法の改善のための実践研究．https://www.mext.go.jp/a_menu/shotou/new-cs/1401806.htm

岡田謙介　2011　心理学と心理測定における信頼性について．教育心理学年報，54，

71-83.