公益財団法人　日本英語検定協会

2017 年度　英語教育研究センター　委託研究

Mixture Rasch Model による英語能力の規準設定

大友賢二・中村洋一・法月健

はじめに

本研究は、2011 年度の財団法人日本英語検定協会英語教育センター委託研究に端を発している。2011 年度から 2013 年度までは「言語テストの規準設定」、2014 年度は「ICT 等を活用した評価についての調査・研究」、2015 年度から現在までは「Mixture Rasch Model による英語能力の規準設定」として研究を継続してきた。6 年間の研究成果については、報告書、あるいは PowerPoint のスライドで、財団法人日本英語検定協会英語教育センターのウェッブ・ページ上で公開している (https://www.eiken.or.jp/center_for_research/contract/)。また、2016 年 3 月 18 日の英語検定協会特別講演会にて、大友賢二研究代表が行った「Mixture Rasch Model による英語能力の規準設定　検討結果と今後の課題」と題した講演でも報告している。

この研究が一貫して検討してきたのは、「規準の設定を客観化するための研究と実践」(大友, 2012, p. 1)である。比較区的広範に及ぶ「言語テストの規準設定」というテーマのもとで研究を始め、Mixture Rasch Model の可能性や課題に焦点を絞りながら、現在の研究を展開している。本項では、まず先行研究を再度振り返りつつ、本研究のこれまでの経過の中から現在の焦点である Mixture Rasch Model の可能性や課題に関連のある議論を概観し、今後の研究における方向性を見定める論点の整理を行う。それに基づき、実際のテストデータを用いて分析を行い、英語能力の規準設定の方法論を検討する。

英語能力の規準設定: ラッシュモデル、潜在ランク理論

『言語テストの規準設定報告書』 (大友 他, 2012, p. 3) は、規準設定の意味と歴史を概観し、Cizek, G. J. (2006, p. 226) の “ … standard setting can be defined as a process by which a standard or cut score is established.” を引き、「規準設定 (standard setting) という用語は、規準、つまり分割点を設定する手順であると規定することができる」とし、この研究の目的が「分割点の設定」であることを明確に示した。

規準設定におけるラッシュモデルの有用性

『言語テストの規準設定報告書』 (大友 他, 2012) の中で、法月 (2012a) は、「規準設定におけるラッシュモデルの有用性」を検討し、ラッシュモデルに基づくシステムは最善で（恐らく唯一の）規準維持の状況を説明する方法である (Bramley, 2010, p. 3)」との指摘を基に、本研究の出発点を設定した。その先行研究のカバーでは、従来の規準設定方法におけるラッシュモデル応用の成果とさらに追求すべき課題をまとめた。その中で、近年の研究で評定者の厳格性・寛容性の差異に関する課題を克服する可能性があるとされる多相ラッシュモデル (Many Facet Rasch Model: MFRM) を取り上げた。また複数の潜在的な母集団を含んだテストデータを分析するために、ラッシュモデルと潜在クラス分析 (Latent Class Analysis: LCA) モデルを統合した Mixture Rasch Model に言及し、Jiao et al. (2011) と Templin & Jiao (2012) などを引き、MRM の可能性に触れた。

規準設定における潜在ランク理論の有用性

ラッシュモデルに続き、法月 (2012b) は Shojima (2007) の Neural Test Theory: NTT の有用性を検討した。NTT は近年では Latent Rank Theory: LRT (潜在ランク理論) と呼ばれ、「学力を順位尺度上で段階評価するためのテスト理論 (荘島, n.d., 小泉・飯村, 2010)」である。「LRT の規準設定における有用性に関する主要な特徴を、古典的テスト理論 (CTT) や項目応答理論(IRT) との比較」をしながら検討した結果、「LRT の分割点設定が、CTT やラッシュモデルを使用する場合とは異なってくる可能性が示唆された」が、「LRT、2&３パラメーターIRT モデル、1 パラメター・ラッシュモデル」のいずれの分析も、「今回の分析データ・条件に関しては一様分布か正規分布の分析を指定することで、間隔尺度とも比較しやすい規準設定が可能になることが確認できた」。さらに、「LRT を使用した規準設定の方向性に示唆を与える近年の言語テストの研究」に触れ、「大きな受験者サンプルを、等化手続きによって同じ潜在尺度上で経年比較することで、規準設定の基盤を築くことが期待される」、「LRT の潜在ランクを使って順序尺度上で最初から段階評価を行うことで、クラスをどこで分けるかの判断が容易にできるようになることと、今後 LRT によって推定される能力（潜在ランク）と発達・習得段階の行動や態度を結びつけて考えることができるようになることに期待を寄せている」、「LRT をラッシュモデルと併用することで、学内で開発された Can-do statements（CDS）の妥当化に効果があったことを指摘」、「LRT は IRT と同じように多値型データにも活用することができる」といった研究の方向性について言及した。そして、今後の研究課題として、「MRM のようなモデルの発展は、LRT の追求する段階評価と他の評価情報との統合の理念から考えて現実的ではないかもしれない」とする懸念に触れながらも、MRM には「複合的な視点から新しい規準設定手続きの構築へと発展していく素地」ができている、ことについても触れている。

法月 (2013) は、「受容語彙を測定するプレイスメントテストにおけるラッシュモデルと潜在ランク理論に基づく規準設定の試行」研究を行い、プレイスメントテストのデータについて「ラッシュモデルと潜在ランク理論を併用することで、合理的な手順で分割点設定を行うことができることが確認でき」、「評定者が多数いなくても、何回も協議を重ねるだけの時間的余裕がなくても、実施することが可能である」との結論を得て、「受験者や項目の分離指標、測定誤差、潜在ランク理論の目標潜在ランク分布と付与されるランクと応答様式の関係等、規準設定の視点から、ラッシュモデルと潜在ランク理論の応用について研究を続けていくことが望まれる」とまとめた。

法月 (2014) は「実用英語検定の級別頻出単語に基づく英語受容語彙力テストの開発と規準設定」のテーマで、ラッシュモデルと潜在ランク理論を用いた規準設定法の検討を行った。その結果について、「本研究の規準設定の手続きは法月 (2013) に基盤を置くものであったが、最終的に採択した方法は、それとはかなり性格の異なるものであった。法月 (2013) では、当該ランクの最も低い領域の正答率が切り替わる地点に分割点を置いたが、本研究では当該ランクの最も高い地点の正答率が切り替わる地点が選ばれた。規準設定の目的にもよるが、受け入れ人数に制限のあるクラス編成や教育プログラムへの入学・参加許可においては、どの方法を選ぶかは実際の受験者のスコア分布に依存する傾向が高いのではないだろうか」と考察した。そして「Zeiky, Perie & Livingston (2008) が主張するように、「分割点は客観的に決めることはできないが、客観的に適用することができるもの」であるならば、２つの統計手法を使って、一定の条件下で分割点設定の方法を客観的に決定した本研究の手法は、相応に評価できるだろう」とまとめた。

以上のような、潜在ランク理論とラッシュモデルを併用した規準設定の方法を検証した研究を経て、2014 年度からは、本研究の 2012 年度の報告書にすでに言及があった Mixture Rasch Model: MRM の可能性に焦点を当てて検討を続けることとした。

Mixture Rasch Model による英語能力の規準設定

大友 他 (2015) で、池田は単純 Rasch モデルが持つ限界に言及し、MRM について「表面的に観ただけではわからないような応答者が持つ潜在特性 (latent trait) や内在的 (internal) 施行法の違いなどからくる偏りに対しては、新しい分析手法を考える必要がある。そうして生まれたのが混合 Rasch モデルである」と紹介している。続いて大友は、Kelderman & Macready (1990), Mislevy & Verhelst (1990), Rost (1990) を引き、「純粋な順序尺度に基づく『潜在ランク理論』(Latent Rank Theory) に連続尺度の精度向上を目指す『項目応答理論』(Item Response Theory) である Rasch Model の特徴を補充することで、より明確な、より実用的な規準設定法を見いだしていこうとする」MRM 研究の方向性を示した。中村は、Jiao et al. (2011) から、"… the proposed mixture Rasch model based method results in a reasonably high level of classification accuracy (p. 514)" を引き、Mixture Rasch model の可能性を示唆した。そ

して、大友 他 (2016) は次項で触れる MRM に関連する統計手法について、文献研究を行った。

ラッシュモデル (Rasch Model: RM)

大友 他 (2016) では、単純ラッシュモデル (Rasch Model: RM) の利点として、「どんな異なったテストを用いても共通の尺度上で能力測定が可能であるということ (test-free person measurement)」、「どんな受験者集団に実施しても、共通の項目特性に関する値を求めることが可能なこと (sample-free item calibration)」、「能力ごとにわかる測定の精度 (multiple reliability estimation)」の 3 点をあげ、「これまでの古典的テスト理論では不可能であったが、それを克服している IRT で実現可能になった」とまとめた。「IRT の特徴の大きな利点は、テストに含まれる項目の難易度とそのテストの受験者の能力を分離して表現できることである」。

潜在クラス分析 (Latent Class Analysis: LCA)

大友 他 (2016) では、Templin & Jiao (2012) を引き、LCA について「観察できるデータから観察できない『潜在特性』を推定するという点で、古典的テスト理論の枠内で行う因子分析と関連がある … (中略) IRT も LCA も、潜在特性モデルという点は同じであるが、IRT は受験者の項目に対する応答に潜在する能力を連続的なパラメータとして推定し、LCA は、潜在的で分割可能なクラスを推定して、受験者の応答を定義する点が異なる」と説明した。また、LCA の利点について、三輪 (2009) の「節約の原理、希薄化の修正」を引き、「たくさんの観測値間の関係を比較的少数の潜在変数へと縮約することでより単純な解釈が得られやすいこと」と「測定誤差を除去した潜在変数間の関係の推定はより精度が高いものになりやすいこと」をあげた。

混合ラッシュモデル (Mixture Rasch Model: MRM)

大友 他 (2016) では、単純ラッシュモデルの分析で前提条件とされている一次元性 (unidimensionality) が確保できないという心理測定的な制約に対応するために検討が進められてきた、Mixture Rasch Model (MRM：混合ラッシュモデル) の先行研究 (Rost & Langeheine, 1994, Cohen, Wollack, Bolt & Mroch, 2002, Kreiner, 2007, Jiao, Lissitz, Macready, Wang & Liang, 2011, Lee & Chen, 2011, Templin & Jiao, 2012, Baghaei & Carstensen, 2013) を検討した。そして、「MRM は複数の潜在的な母集団を含んだテストデータを分析するため、ラッシュモデルと潜在クラス分析 (latent class analysis: LCA) モデルを統合したモデルであり、テストデータと主観的な審査員の判定を融合した規準設定手続きを導くこともできる (Templin& Jiao, 2012, p. 387, p. 379) が、通常は審査員の判定を伴わずに実施することができる (Lissitzi, 2013, p. 170) 統計的解決モデルと言える」と考え、Jiao et al. (2011) の分析手

続きに着目することとなった。

法月 (2016) は、法月 (2014) と同一のデータを使って、２～５クラスの分析を行い、得られた結果を比較検討した。Jiao et al. (2011, pp. 520-522) の示した計算手続きに沿ってクラス間の分割点を求めると、５クラス分析においては Class 3 と Class 4、Class 4 と Class 5 のそれぞれの平均値間に、図１のように交点が確認されたが、Class 1 と Class 2、Class 2 と Class 3 については、図２のように適切な分割点は得られなかった。

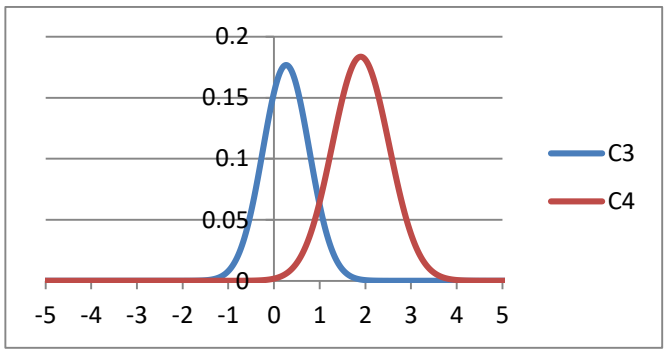

図 1. ５クラス分析における C3 と C4 の正規分布曲線

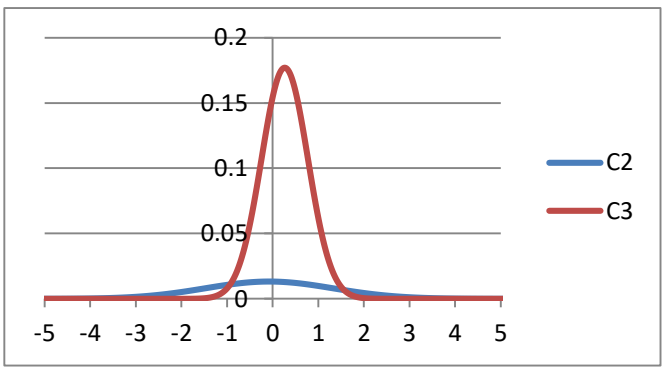

図 2. ５クラス分析における C2 と C3 の正規分布曲線

研究課題

法月 (2017) は、213 名の受験者に実施した英語語彙テスト (50 問) に対して、Jiao et al. (2011) の MRM 分析に基づき分割点設定を行った結果、2クラス、3クラス分析では、いずれのクラス間でも分割点を設定することができたが、4クラス、5クラス分析では、いずれも複数の隣接するクラス・ペア間において分割点が設定できない状況を確認している。

MRM を使った分割点設定の研究事例はそれほど多いとは言えないが、Jiao et al. (2011)は 10,000 人のシミュレーションデータで5クラス分析、Kreiner (2007)や Kreiner et al. (2006) は 1,000 名前後の 2 クラス分析の結果を中心に議論を行っている。特に分割するクラス数が多い場合は、受験者のサンプルサイズがかなり大きくないと有効な結果が得られないのだろうか。

法月 (2017) は、193 名の英語語彙テストの4段階の理解度自己評価データ (50 問) を使って、MRM 分析に基づく分割点設定について検証を行ったが、2クラスから5クラス分析のいずれのクラス間においても有効な分割点設定ができなかったことを示している。異なる種類の多値型データに対して分析を行った場合、MRM による有効な分割点設定は可能だろうか。

Jiao et al. (2011) は、LCA 分析では、評定者の主観的な判断に頼ることなく、テストデータに適合するモデルを軸にして受験者を分類できるが、MRM 分析と異なり、潜在グループの分類と分割点の設定を、2段階で行わなければならないとしている。LCA と単純ラッシュモデル(RM) の分析 を2段階で実施した場合、MRM を行った場合と同じ結果が得られるであろうか。

法月 (2017) は、213 名の英語語彙テスト2値型データと 193 名の理解度自己評価多値型データに対して潜在ランク理論（LRT）と RM の分析を2段階で実施した場合、2ランクから5ランク分析のいずれのクラス間でも分割点を設定することができたことを示した。異なる種類の2値型、多値型データでも同様の結果が得られるであろうか。

本研究報告では、以上の観点に基づき、以下の研究課題を基軸に分析を行い、その結果に基づき、議論を行うものとする。

研究課題1： 分析データの受験者数が増えるにつれて、MRM による分割点設定の効果が高まると言えるだろうか。

研究課題2： 多値型データにおいても、MRM 分析を行うことで、分割点設定ができるだろうか。

研究課題3： 単純 RM と LCA を組み合わせた RM-LC 法による分割点設定においても、MRM と同様の効果を得ることが可能だろうか。

研究課題4： 単純 RM と LRT を組み合わせた RM-LRT 法による分割点設定においても、MRM と同様の効果を得ることが可能だろうか。

分析方法

被験者

本研究の実施に当たって、日本英語検定協会英語教育研究センターより、20,000 人の受験者で構成される英語能力試験のデータ （ELP20,000） 提供を受けた。このデータから、MRM 分析におけるサンプルサイズの影響を比較分析するために、様々な大きさの受験者集団を無作為に抽出し、技能別に全データの 20,000 人 (ELP20,000)、その一部の 10,000 人 (ELP10,000)、5,000 人 (ELP5,000)、1,000 人 (ELP1,000)、500 人 (ELP500)、250 人 (ELP250)、100 人 (ELP100) のサンプルから形成されるデータを作成した。

分析に用いたテスト

分析した ELP20,000 は、リーディング (R) 部門が 41 問、リスニング (L) 部門が 29 問で構成され、各問 4 肢択一式の正解 1 点、不正解 0 点の2値型データの問題である。20,000 人の全デ

ータから、10,000、5,000、1,000、500、250、100 人のサンプルサイズの2値型データを、L、R の技能部門別に、無作為に抽出して、分析を行った。さらに、L 部門のうち、5つの説明文に対応するペア項目群への 20,000 人の解答結果を、2 問不正解 0 点、1 問正解 1 点、2 問正解 2 点の多値（polytomous）型データ (L20,000_P) として分析し、そこから、異なるサイズの多値型データ分析用のサンプル (L10,000_P、L5,000_P、L1,000_P、L500_P、L250_P、L100_P) を無作為に抽出して、比較を行った。

表 1 は、受験者 20,000 人の、(R) + (L) 70 問、(R) 41 問、(L) 29 問の3つのデータセットについて、Test Data Analysis Program: TDAP Ver. 2.02 (大友・中村・秋山, 2008) を使用して算出した基礎統計量、信頼性係数 (クロンバックの α)、項目分析の結果一覧である。

表1

基礎統計量・信頼性係数・項目分析

	(R) + (L) 70 問	(R) 41 問	(L) 29 問
Minimum score	0	0	0
Maximum score	69	41	29
Median	34	20	14
Range	69	41	29
Mean	34.255	20.199	14.056
average proportion of passing	0.489	0.493	0.485
Variance	76.383	38.304	19.412
Standard deviation	8.740	6.189	4.406
Skewness	0.234	0.159	0.222
Kurtosis	-0.357	-0.496	-0.230
Coefficient Alpha	0.811	0.776	0.714
Discrimination Power Index (Average)	0.266	0.317	0.326
Actual equivalent number of options (Average)	3.165	3.193	3.126

3つのデータセットとも、平均正答率は 0.500 弱で、歪度は、0.234、0.159、0.222 でやや右よりの分布、尖度は -0.357、-0.496、-0.230 と平坦な分布状態を示している。信頼性係数は 0.811、0.776、0.714 であった。

項目分析の結果としては、弁別力指数の平均が 0.266、0.317、0.326 と算出された。また実質選択肢数の平均は、3.165、3.193、3.126 と比較的高い数値であった。

このような古典的テスト理論の分析から得た、平坦な分布状況や信頼性係数の値から鑑みると、素点の処理だけでは、能力の幅が広い受験者が受験するこのテストの規準設定に関する情報

が不足するのではないかとの印象がある。また、項目分析から得た弁別力指数や実質選択肢数の値から、20,000 人の受験者の中に、いくつかの潜在的な特性を持つグループが存在し、それぞれ異なる解答ストラテジーを持つのではないか、ということが推察される。

分析手法

WINMIRA 2001 (von Davier, 2001) に基づき、2～5クラス MRM 分析を行い、Jiao et al. (2011)、法月 (2016) が説明する方法で確率密度関数を計算した。隣接するクラス Cx と Cx+1 の両平均値の間に交点が位置する場合は、妥当な分割点と判定し、交点の位置が両平均値の間に位置しなかったり、交点が得られない場合は、当該クラス間で妥当な分割点が存在しないと判定することとした。WINMIRA 2001 の分析は、初期設定の条件で、異なるサンプルサイズを分析したが、一部の分析については、各潜在クラスに帰属する受験者の割合を事前に設定して行った分析結果と比較した。

RM と LC を2段階で行う RM-LC 分析についても、WINMIRA 2001 を使用したが、RM と LRT を 2 段階で行う RM-LR 分析については、LRT の分析部分は、Exametrika (荘島, 2011) を使用した。RM-LC 分析、LR-LR 分析とも、分割点を求めるための確率密度関数の計算は、MRM 分析と同じ方法で行った。

実験計画

研究課題1： 受験者サンプルサイズと分割点設定の効果

まず、ELP20,000 の受験者データを正解項目 1 点、不正解項目 0 点の2値型データに変換した。次に、データをリーディング部門とリスニング部門に分け、それぞれの部門の全体のデータから 10,000 人、5,000 人、1,000 人、500 人、250 人、100 人を無作為抽出した。

それぞれのサンプルの2値型データを WINMIRA 2001 の分析ファイルに入力し、初期設定の状態で、MRM 分析を行った。次に、MRM 分析によって得られた各潜在クラスの能力平均値が昇順になるように、クラスの順番を並べ替え、受験者数、各クラスに所属する受験者の比率、各クラスの平均値、標準偏差を集計し、それらの数値を Jiao et al. (2011) が提案した下記の数式に代入し、隣接するクラス間で確率密度曲線が交わる地点となる分割点を計算した。
(Jiao et al., 2011, pp.520-522; 大友・中村・法月、2016, pp.37-38 を参照)。

$$W_1 \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} = W_2 \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}} \quad (1)$$

W_I: 平均能力値が低い方から順にクラスを並べ替えたときのあるクラス$_{(I)}$に所属する受験者数の全体における割合

W_2: あるクラス$_{(I)}$に隣接するより平均値のより高いクラス$_{(II)}$に所属する受験者数の全体における割合

x: 隣接するクラスの正規分布曲線が交差する地点

μ_1:クラス$_{(I)}$の得点の平均

μ_2:クラス$_{(II)}$の得点の平均

σ_1:平均値がより低いクラス$_{(I)}$の得点の標準偏差

σ_2:平均値がより高いクラス$_{(II)}$の得点の標準偏差

e: 定数 e は自然対数の底で、2.71828182845904 となる。

(例) $e^2 = 2.71828182845904^2 = 7.389056099$

法月 (2017) は、213 名の受験者に実施した英語語彙テスト (VKS213) が、潜在4、5クラスの MRM 分析において、隣接するクラス間に適切な分割点が見られない結果を多く生じたことを示している。たとえば、VKS213 の5クラス MRM 分析において、潜在クラスの受験者平均能力値を低い順に C1、C2、... C5 と並べ替えると、C1の能力平均値が -0.78、C2 の能力平均値が -0.07 であったが、C1/C2 の確率密度曲線の交点を計算すると、-0.89 となり、両クラスの平均値の間に位置しない値を示したため、適切な分割点が得られなかった。同様に、C2/C3、C3/C4 も適切な交点が得られなったが、C4/C5 は両クラスの平均値の間に位置する妥当な交点を確認することができた。各潜在クラスの所属受験者数は、C1 から 16 名、48 名、10 名、 61 名、78 名だったが、適切な分割点が得られなかったクラス間は、いずれも所属受験者数が全データの5%前後しかない C1 や C3 が関係している。

本研究では、20,000 人の受験者に実施した ELP20,000 のデータをリーディング(R) 部門、リスニング (L) 部門に分け、各部門の全データから、10,000 人、5,000 人、500 人、250 人、100 人で構成されるサンプルを抽出し、サンプルサイズ（ELP10,000、ELP5,000、ELP500、ELP250、ELP100）が大きくなるにつれて、分割点設定の効果が上がると言えるかどうか、技能部門別に比較することとする。

WINMIRA 2001 では、初期設定を変更して、各クラスに所属する確率を事前に設定することもできる。VKS213 データの5、4クラス分析のように各クラスに所属する受験者の分布がかなり偏る場合は、この機能を使うことで、偏りを是正し、分割点設定の効果が向上することが可能かどうかについても、検証することとする。

研究課題2： 多値型データの分割点設定

VKS テストにおいては、奇数番号のテスト項目の直後に位置する偶数番号の「回答欄」に、その「項目の単語」に対する理解度を4段階（5－かなり知っている単語　4－何となく意味がわかる単語　2－見たことはあるが意味は分からない単語　1－見たこともないし、意味もわからない単語）で、受験者に自己評価させている (法月、2014)。法月 (2017) は、この理解度自己評価

への回答者 193 名の多値型 (polytomous) データ (VKS193_P) に対して、WINMIRA 2001 を使用して、MRM 分析を行ったが、2～5クラス分析のいずれのクラス間においても適切な分割点は得られず、交点すら存在しない状況が大半を占める結果となった。

VKS193_P の MRM 分析では、サンプルサイズが、VKS の2値型データよりもさらに小さいだけでなく、クラスサイズの偏りが非常に大きくなり、5クラス、4クラス分析では、データ全体の3%、10%に満たないクラスが1つずつ生じ、3クラス分析ではデータ全体の 10%未満のクラスが一つあり、2クラス分析でも2クラスの比率が約2対8の割合となったことも、分割点設定を困難にした要因と言えるかもしれない。

本研究において扱う ELP20,000 は、R 部門、L 部門の項目ともに、いずれも元来は2値型データであるが、L 部門の5つの説明文に対応するペア項目群への解答結果を、各ペア 2 問とも不正解の場合は 0 点、1 問正解の場合は 1 点、2 問とも正解の場合は 2 点と計算し、多値型データ (L20,000_P) として分析する。ELP20,000_P は VKS193_P と比べて、項目数は少ないが、受験者数が大幅に多く、「テストデータ」でもある。

また、2値型データの分析の時と同じように、10,000 人、5,000 人、1,000 人、500 人、250 人、100 人のデータ（L10,000_P、L5,000_P、L1,000_P、L500_P、L250_P、L100_P）を抽出して、サンプルサイズの影響を比較し、クラスサイズに偏りが出た場合は、事前にクラスサイズの確率を設定して、分割点設定の効果を探ることにした。

研究課題3： RM-LC 法による分割点設定

VKS213 や VKS193_P のデータについて、これまでに単純ラッシュモデル(RM) と潜在クラス分析 (LCA) を個別に行った後に、確率密度曲線を計算して、分割点設定を行う方法 (RM-LC 法) が実践可能か否かは、報告されていない。国際的な数学テストを MRM と LCA を使って分析した Toker (2016) は、それらの分析結果に顕著な差異を確認している。もし、ラッシュモデルと潜在クラス分析を融合した方法に基づく分割点設定においても、MRM 分析に基づくものと大きく異なる結果をもたらすならば、この RM-LC 法の効果について検証する価値があると言える。そこで、2値型の VKS213 データを使って、RM-LC 法で分析したところ、MRM 分析では機能しなかった4、5クラス分析も含めて、実施した2から5クラスの分析のすべての隣接クラス間に適切な分割点を設定できることを確認した。他のテストデータや多値型データにも同様の効果が得られるのか、2値型の ELP20,000、ELP10,000、ELP5,000、ELP1,000、ELP500、ELP250、ELP100 を R 部門、L 部門についてそれぞれ比較し、多値型の L20,000_P、L10,000_P、L5,000_P、L1,000_P、L500_P、L250_P、L100_P についても、MRM 分析と同様の検証を行う。

本研究においては、RM 分析、LCA 分析ともに、WINMIRA 2001 を使用する。

研究課題4： RM-LR 法による分割点設定

我々のこれまでの研究の中で最も安定した分割点設定を示してきたのが、 単純ラッシュモデ

ル(RM) と潜在ランク理論 (LRT) の分析を個別に行った後に、確率密度曲線を計算して、分割点設定を行う RM-LR 法であった。法月（2017）は、VKS213 の2値型データ、 VKS193_P の多値型データの2から5クラスの実施したすべての分析において、適切な分割点を得ることができたことを示しているが、MRM 分析で、多値型分析がまったく機能しなかったのとは対照的な結果であった。MRM が機能しなかった要因が、サンプルサイズが小さく、分析に適さなかったためだとしたら、様々なサイズの ELP テストを異なる方法で分析する価値は高いと言える。RM-LR 法の LRT の分析において、法月 (2017) は、VKS の受験者数が少なかったため、植野・荘島 (2010, p.98) が「毎回の計算が必ず一致し、また計算速度が速い」方法として説明している生成トポグラフィックマッピング(GTM) による推定法を使用することができなかった。ELP テストデータを使って、GTM と VKS データ分析でも使用してきた自己組織化マッピング（SOM）による推定法を比較することで、RM-LR 法の汎用性を探ることも期待される。

分析結果

研究課題1： MRM 分析における受験者サンプルサイズと分割点設定の効果

まず、20,000 人の受験者サンプルの2値型解答データに基づく、ELP20,000 のリスニング部門（L）（29 問）、リーディング部門（R）（41 問）の分割点設定の結果を、表2と表3を使って、説明することにする。

表2で、2値型の L 部門の5クラス分析 (5C) について、C1と C21 の横の数値を見ていくと、元々の分析で得られた潜在クラスを能力平均値昇順で並べ替えて、最も平均値の低い C1 に配置された受験者数は 4,224 人で全体比 0.21、次に平均値の低い C2の受験者数は 3,557 人で全体比0.18だった。平均値（$\bar{x}$）はそれぞれ -0.93 と -0.25 で、分割点（CP）の下の -0.55 はC1と C2 クラスの確率密度曲線が交わる地点の数値を示している。この値が隣接する両クラスの平均値の間に位置する値であることから、適切な分割は「可」と解釈した。その一方で、同じ5クラスの C2 と C3 の確率密度曲線の交点は-1.15 を示したが、この値は両クラスの平均値の間に位置していないため、妥当な分割点とは解釈されない。このように、5クラス分析については、すべての隣接するクラス間で分割の条件を満たすには至らかったため、表の右欄の規準設定後の新しいグループの人数に関する情報は空欄にしてある。

一方、4、3、2クラス分析では、いずれも隣接するクラス間で妥当な分割点が得られたことが示されていて、右欄には規準設定後の新しいグループの人数に関する情報が提示されている。MRM 分析では、単純ラッシュ分析と異なり、受験者パラメター値と素点との関係が一律ではないが、両得点指標を提示した受験者得点一覧を、受験者パラメター値が昇順もしくは降順になるように並べ替えると、例えば4クラス分析 C1/C2 の分割点 -0.66 が位置する地点は、素点の 10 点と 11 点の間に収束していると解釈できる。C1 の分割後人数の 4,460 人のうち、もともと C1 に位置付けられていた受験者の数が 3,633 名、C2 が 816 名、C3 が 10 名、C4 が1名いることがわかり、もともとのクラス編成とは少し異なる構成員になっていると言える。

表2
MRM 20,000 人リスニング・テスト(29 問)・データ分析結果

		人数	比率	$\bar{x}$	SD	CP	可否	CRS	NSS	OC1	OC2	OC3	OC4	OC5
5C	C1	4224	.21	-0.93	0.43	-0.55	○							
	C2	3557	.18	-0.24	0.39	-1.15	×							
	C3	4879	.25	-0.23	0.37	0.16	○							
	C4	4723	.24	0.54	0.39	1.15	○							
	C5	2617	.13	1.42	0.71	—	—							
4C	C1	4766	.24	-0.89	0.43	-0.66	○	10/11	4460	3633	816	10	1	—
	C2	6623	.33	-0.24	0.36	0.12	○	15	7330	1133	5027	1133	37	—
	C3	4888	.24	0.31	0.38	0.86	○	19	5009	0	772	3223	1014	—
	C4	3723	.19	1.25	0.68	—	—	—	3201	0	8	522	2671	—
3C	C1	5965	.30	-0.82	0.44	-0.51	○	11/12	6009	4909	1100	0	—	—
	C2	9939	.50	0.01	0.41	0.75	○	18/19	10764	1056	8465	1243	—	—
	C3	4096	.20	1.21	0.66	—	—	—	3227	0	374	2853	—	—
2C	C1	10284	.51	-0.57	0.51	0.03	○	14/15	11073	9227	1846	—	—	—
	C2	9716	.49	0.59	0.69	—	—	—	8927	1057	7870	—	—	—

*CP:分割点,、可否: 適切な分割点設定の可否、CRS: 分割点(素点)、
NSS: 規準設定後の新しいクラス(グループ)の人数、OC1～5:規準設定前の所属クラス

表2の結果から、サンプルサイズに関連して特筆すべきことは、213 名の VKS データでは3クラス分析までしか機能しなかったが、ELP20,000 の L 部門では、4クラスまで分析できたことであろう。

次に表3を見ると、R 部門では、適切な交点が得られないクラス・ペア数が増え、条件を完全に満たしているのは、2クラス分析のみである。この結果からは、単純に分析する受験者数が多くても、無条件でデータが分析に適応するわけではないことが、明らかである。

表3
MRM 20,000 人リーディング・テスト(41 問)・データ分析結果

		人数	比率	$\bar{x}$	SD	CP	可否	CRS	NSS	OC1	OC2	OC3	OC4	OC5
5C	C1	4955	.25	-0.79	0.38	-0.53	○							
	C2	6348	.32	-0.16	0.38	0.63	×							
	C3	1738	.09	0.08	0.53	-0.13	×							
	C4	3393	.17	0.52	0.53	0.71	○							
	C5	3566	.18	0.94	0.54	—	—							
4C	C1	5298	.26	-0.77	0.38	-0.51	○							
	C2	7608	.38	-0.08	0.41	0.48	×							
	C3	3857	.19	0.46	0.53	0.83	○							
	C4	3237	.16	1.00	0.53	—	—							
3C	C1	9301	.47	-0.57	0.44	0.02	○							
	C2	7083	.35	0.48	0.59	×*	×							
	C3	3616	.18	0.57	0.55	—	—							
2C	C1	9959	.50	-0.54	0.45	-0.01	○	20/21	10591	8898	1693	—	—	—
	C2	10041	.50	0.54	0.56	—	—	—	9409	1061	8348	—	—	—

*CP 欄の×は交点なし

表4と表5は、ELP20,000 から無作為抽出された受験者のデータを使って、異なるサンプルサイズの分析結果を比較したものである。全体的な傾向として、サンプルのサイズが小さくなると、結果的に、分割点設定が難しくなっている状況が確認できる。

表4
MRM リスニング・テストの分析結果―サンプルサイズによる影響

	ELP20,000 (L)				ELP10,000 (L)				ELP5,000 (L)				ELP1,000 (L)			
	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5
2C	○				○				○				○			
3C	○	○			○	○			○	○			○	○		
4C	○	○	○		○	×	○		○	×	○		×	○	○	
5C	○	×	○	○	○	×	○	○	○	×	○	○	×	×	○	○

	ELP500 (L)				ELP250 (L)				ELP100 (L)			
	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5
2C	○				○				○			
3C	○	○			○	○			×	○		
4C	×	○	○		○	○	×		×	×	×	
5C	○	×	×	×	×	○	×	○	×	×	×	○

表5
MRM リーディング・テストの分析結果―サンプルサイズによる影響

	ELP20,000 (R)				ELP10,000 (R)				ELP5,000 (R)				ELP1,000 (R)			
	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5
2C	○				○				○				○			
3C	○	×			○	×			○	×			○	×		
4C	○	×	○		○	×	○		×	×	×		×	○	×	
5C	○	×	×	○	○	×	×	○	○	×	×	×	○	×	×	×

	ELP500 (R)				ELP250 (R)				ELP100 (R)			
	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5
2C	○				○				○			
3C	○	×			○	○			×	×		
4C	×	○	×		×	×	×		×	×	×	
5C	×	×	×	×	×	×	×	×	×	×	○	×

VKS の 213 名2値型データでは、2、3クラス分析を実施することができた。受験者サンプルは ELP20,000、10,000、5,000、500 に比べてかなり小さかったが、問題形式がすべて類義の単語を選ぶ趣旨の単一種類の問題であったことや、問題数も 50 問と少なくなかったことは、規準設定の観点からは、MRM の分析に比較的適していたのかもしれない。しかしながら、同じ問題であれば、サンプルサイズが大きいほうが分析において安定感が得られることについても、今回扱った2値型データの分析から、おおむね確認できたように思える。

研究課題2　MRM による多値型データの分割点設定

表6は上述のように、同じ説明文(リスニング)の内容に関する2問ずつの問題をセットにして、多値型データとして分析し、得られた L20,000_P の分割点設定の結果をまとめたものである。

表6
MRM 20,000 人リスニング・テスト・多値型データ(6問 12 満点)分析結果

		人数	比率	$\bar{x}$	SD	CP	可否	CRS	NSS	OC1	OC2	OC3	OC4	OC5
5C	C1	4891	.24	-1.20	0.70	-1.41	×							
	C2	4557	.23	-1.04	0.53	-0.44	×							
	C3	3153	.16	-0.84	0.51	-0.55	○							
	C4	6369	.32	0.09	0.45	1.07	×							
	C5	1030	.05	0.92	0.80	—	—							
4C	C1	8200	.41	-1.18	0.63	-0.78	○	4	8520	6873	1007	640	0	—
	C2	4407	.22	-0.71	0.41	-0.43	○	4/5	1858	0	1858	0	0	—
	C3	6502	.33	0.03	0.45	1.02	○	8/9	9264	1327	1542	5858	537	—
	C4	891	.04	1.11	0.71	—	—		358	0	0	4	354	—
3C	C1	7011	.35	-1.11	0.69	-1.20	×							
	C2	10562	.53	-0.58	0.55	0.43	○							
	C3	2427	.12	0.73	0.59	—	—							
2C	C1	16768	.84	-0.82	0.63	0.33	○	6/7	17567	16600	967	—	—	—
	C2	3232	.16	0.62	0.55	—	—	—	2433	168	2265	—	—	—

クラス分布が2値型データでは均等化する傾向がある2クラス分析でさえも、一方のクラスが全体の 84%近くを占めてしまうなど、偏りが目立ち、確率密度関数を計算するまで、分割できるかどうかがわからない「変則性」は強く感じられたが、VKS193_P の 193 名のデータでは全く機能しなかった MRM 分析による分割点設定が、一応は、2クラスと4クラス分析において成立する結果となった。

しかしながら、実践的な解釈は必ずしも容易ではない。4クラス分析では、C2/C3 の分割点は素点に換算すると4点と5点の間に位置する結果となったが、小数点の下位区分がないため、実質的には、C1/C2 と同じ4点となる。ちなみに、分割点設定後の C2 の 1,858 人は、全員、旧 C2 に所属し、-0.780 の受験者パラメター値が付与され、素点は4点だった。一方、分割点設定後の C1 の 8,520 人のうち 2,797 人の素点も4点だったが、そのうち 640 人の受験者パラメター値が -0.789 で、旧 C1 に所属し、残りの 2,157 人の受験者パラメター値は-0.792 で、旧 C3 に所属していた。MRM の対数値ではわずかながらも差があって、合理的な分割点形成に至っているが、規準設定の情報を必要としたり、その影響を受ける人たちに対しては、同一の素点の受験者がなぜ異なるクラスに分類されるのか、わかりやすく説明することが求められるだろう。

多値型データについても他のサンプルサイズと比較を行ったところ、表7のような結果となった。「○」の数では L10,000_P データが L20,000_P よりも多く、「×」の数では、L1,000_P が、L500_P や L100_P よりも圧倒的に多くなったが、L500_P の5クラス分析で C4、C5、L100_P の4クラス分析で C4、5クラス分析で C5 の所属者が 0 人、C1 は 1 人となり、これらのクラスについては、確率密度曲線の計算ができない。分割点設定の成否については、状況が予測不能な部分はあるが、比較的小さなサンプルサイズにおいては、クラス数が増えるほど、成員の確保が難しくなっていくのは間違いないようだ。

表7
MRM リスニング・テスト(多値型データ)の分析結果―サンプルサイズによる影響

	L20,000_P				L10,000_P				L5,000_P				L1,000_P			
	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5
2C	○				○				×				×			
3C	×	○			○	○			○	×			×	×		
4C	○	○	○		×	○	○		×	○	×		×	○	×	
5C	×	×	○	×	×	○	○	○	×	×	○	×	×	×	×	×

	L500_P				L250_P				L100_P			
	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5
2C	○				○				×			
3C	○	×			○	○			×	○		
4C	×	×	○		○	×	×		×	○	なし	
5C	×	○	なし	なし	○	×	×	なし	1名	○	×	なし

＊なし:上位側クラス (C3/C4 の場合は C4、C4/C5 の場合は C5) の所属者なし
＊1名:C1 クラス所属者1名

研究課題3 RM-LC 法による分割点設定

表8は、ELP20,000 の L 部門のデータについて、単純ラッシュモデルと潜在ランク理論を融合した RM-LC 法によって、分割点設定を試みた結果をまとめたものである。MRM の受験者パラメター値ではなく、素点と常に一対一の関係にある単純ラッシュモデルに基づくものを使用した。VKS213 データの RM-LC 法による分析と同様に、ELP20,000 の L 部門では、2～5クラス分析のすべての隣接クラス間で適切な分割点を得られた。

表8
RM-LC 法 20,000 人リスニング・テスト(29 問)・データ分析結果

		人数	比率	$\bar{x}$	SD	CP	可否	CRS	NSS	OC1	OC2	OC3	OC4	OC5
5C	C1	3120	.16	-1.15	0.47	-0.95	○	8/9	2031	1793	238	0	0	0
	C2	8156	.41	-0.38	0.35	0.03	○	14/15	9042	1326	6927	789	0	0
	C3	6157	.31	0.36	0.30	0.86	○	19/20	6551	1	990	5049	511	0
	C4	2268	.11	1.12	0.32	1.89	○	24/25	2142	0	1	319	1753	69
	C5	299	.01	2.30	0.53	—	—	—	234	0	0	0	4	230
4C	C1	4835	.22	-1.04	0.45	-0.74	○	10/11	4459	3598	861	0	0	—
	C2	9187	.46	-0.20	0.35	0.29	○	16/17	9798	787	7848	1163	0	—
	C3	5396	.27	0.63	0.33	1.33	○	21/22	4721	0	478	4100	143	—
	C4	1032	.05	1.70	0.51	—	—	—	1022	0	0	133	889	—
3C	C1	7171	.36	-0.84	0.46	-0.40	○	12/13	7653	6456	1197	0	—	—
	C2	9835	.49	0.11	0.36	0.80	○	19/20	9971	715	8521	735	—	—
	C3	2994	.15	1.20	0.50	—	—	—	2376	0	117	2259	—	—
2C	C1	11494	.57	-0.59	0.51	0.09	○	14/15	11073	10528	545	—	—	—
	C2	8506	.43	0.64	0.55	—	—	—	8927	966	7961	—	—	—

一方、表9のように、R 部門においては、MRM 分析と同様に、5、4クラス分析では適切な分割点が得られないクラス・ペアもあったが、3、2クラス分析ではすべて機能した。

表9
RM-LC 法 20,000 人リーディング・テスト(41 問)・データ分析結果

		人数	比率	$\bar{x}$	SD	CP	可否	CRS	NSS	OC1	OC2	OC3	OC4	OC5
5C	C1	3737	.19	-0.97	0.35	-0.73	○							
	C2	7397	.37	-0.30	0.28	0.15	○							
	C3	3085	.15	0.40	0.32	0.11	×							
	C4	4039	.20	0.46	0.27	0.95	○							
	C5	1742	.09	1.30	0.40	—	—							
4C	C1	7190	.36	-0.75	0.37	-0.32	○							
	C2	6846	.34	0.07	0.30	0.50	×							
	C3	2850	.14	0.39	0.34	0.72	○							
	C4	3114	.16	1.06	0.42	—	—							
3C	C1	6760	.34	-0.78	0.36	-0.37	○	17/18	7189	6315	874	0	—	—
	C2	9064	.45	0.08	0.29	0.59	○	25/26	8582	445	7748	389	—	—
	C3	4176	.21	0.97	0.41	—	—	—	4229	0	442	3787	—	—
2C	C1	10933	.55	-0.55	0.42	0.05	○	20/21	10591	10283	308	—	—	—
	C2	9067	.45	0.60	0.46	—	—	—	9409	650	8759	—	—	—

表 10 と表 11 は、RM-LC 分析のサンプルサイズ別の分析結果の相違を示している。L 部門では、サンプルサイズが小さくなると、徐々に分割点設定が困難になる状況が示されているが、R 部門については、変動は見られるものの、あまりサンプルサイズによる影響は感じられない。

表 10
RM-LC 法 リスニング・テストの分析結果―サンプルサイズによる影響

	ELP20,000 (L)				ELP10,000 (L)				ELP5,000 (L)				ELP1,000 (L)			
	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5
2C	○				○				○				○			
3C	○	○			○	○			○	○			○	○		
4C	○	○	○		○	○	○		○	○	○		○	×	○	
5C	○	○	○	○	○	×	○	○	○	×	○	○	×	○	×	○

	ELP500 (L)				ELP250 (L)				ELP100 (L)			
	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5
2C	○				○				○			
3C	○	○			○	○			×	○		
4C	○	○	○		○	×	○		×	○	○	
5C	○	×	×	○	○	×	×	○	×	○	○	○

表11
RM-LC 法 リーディング・テストの分析結果―サンプルサイズによる影響

	ELP20,000 (R)				ELP10,000 (R)				ELP5,000 (R)				ELP1,000 (R)			
	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5
2C	○				○				○				○			
3C	○	○			○	○			○	○			○	○		
4C	○	×	○		○	×	○		○	×	○		○	○	○	
5C	○	○	×	○	○	○	×	○	○	×	×	○	○	○	×	○

	ELP500 (R)				ELP250 (R)				ELP100 (R)			
	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5
2C	○				○				○			
3C	○	○			○	○			○	○		
4C	○	○	○		○	×	○		○	×	○	
5C	×	○	×	○	○	×	×	○	○	×	×	○

RM-LC 法を L20,000_P の多値型データに適用すると、表12 で示す結果となった。2、3クラス分析では、分割点設定がすべての隣接クラス間で機能したが、2値型データ分析の結果に比べて、両クラスの分析ともクラス分布が大きく偏っている。

表 12
RM-LC 法 20,000 人リスニング・テスト(多値型データ分析)結果

		人数	比率	$\bar{x}$	SD	CP	可否	CRS	NSS	OC1	OC2	OC3	OC4	OC5
5C	C1	4193	.21	-1.45	0.66	-1.51	×							
	C2	4343	.22	-1.09	0.44	-0.65	○							
	C3	4024	.20	-0.25	0.47	×	×							
	C4	7125	.36	-0.16	0.48	1.16	○							
	C5	315	.02	1.72	0.63	—	—							
4C	C1	7625	.36	-1.34	0.59	-0.88	○							
	C2	7821	.39	-0.36	0.52	0.39	×							
	C3	4556	.23	-0.12	0.45	1.05	○							
	C4	358	.02	1.67	0.61	—	—							
3C	C1	7881	.39	-1.27	0.62	-0.90	○	3/4	5723	4981	742	0	—	—
	C2	11797	.59	-0.26	0.52	1.23	○	8/9	13919	2900	11012	7	—	—
	C3	322	.02	1.70	0.63	—	—	—	358	0	43	315	—	—
2C	C1	17836	.89	-0.79	0.67	0.53	○	7/8	19058	17830	1228	—	—	—
	C2	2164	.11	0.71	0.53	—	—	—	942	6	936	—	—	—

RM-LC 法を使って、他のサンプルサイズのデータを分析すると、表 13 の結果となった。MRM 分析と比べると、L1,000_P、L250_P のような例外はあるものの、比較的少ないクラス数がかかわる分析は、安定して機能しているようだ。変動は見られるものの、サンプルサイズによる大きな影響はあまり感じられない。

表 13
RM-LC 法 リスニング・テスト(多値型データ)の分析結果―サンプルサイズによる影響

	L20,000_P				L10,000_P				L5,000_P				L1,000_P			
	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5
2C	○				○				○				○			
3C	○	○			○	○			○	○			×	○		
4C	○	×	○		○	×	○		○	×	○		×	○	○	
5C	×	○	×	○	○	×	×	○	○	○	×	○	×	×	○	○

	L500_P				L250_P				L100_P			
	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5
2C	○				○				○			
3C	○	○			×	×			○	○		
4C	○	×	×		×	×	×		×	×	○	
5C	×	×	×	×	○	×	×	×	○	×	×	○

研究課題4: RM-LR 法による分割点設定

表14 と表15 は、ELP20,000 の L 部門と R 部門の2値型データを GTM による推定に基づき、RM-LR 分析した結果を示している。使用した統計ソフトの初期設定が、「一様分布」を初期設定とする SOM と異なり、GTM では分布を「指定しない」状態であるうえに、MRM や RM-LC 法の分析と類似の条件で比較ができるように、「分布指定なし」の分析をまず行った。

表 14
RM-LR 法（GTM 分布指定なし）20,000 人リスニング・テスト・データ分析結果

		人数	比率	$\bar{x}$	SD	CP	可否	CRS	NSS	OR1	OR2	OR3	OR4	OR5
5R	R1	4330	.22	-1.05	0.44	-0.76	○	10/11	4459	3643	791	25	0	0
	R2	3663	.18	-0.48	0.25	-0.31	○	12/13	3194	631	1886	674	3	0
	R3	4012	.20	-0.11	0.23	0.10	○	15/16	5072	56	964	2916	1135	1
	R4	4109	.21	0.31	0.22	0.66	○	18/19	4048	0	22	397	2813	816
	R5	3886	.19	1.07	0.51	—	—	—	3226	0	0	0	158	3069
4R	R1	5147	.26	-0.99	0.45	-0.66	○	10/11	4459	4011	448	0	0	—
	R2	4885	.24	-0.34	0.25	-0.10	○	13/14	4922	1101	3336	485	0	—
	R3	5249	.26	0.16	0.24	0.54	○	17/18	6236	35	1099	4427	675	—
	R4	4719	.24	0.97	0.51	—	—	—	4383	0	2	337	4044	—
3R	R1	6755	.34	-0.87	0.45	-0.46	○	11/12	6009	5511	498	0	—	—
	R2	7170	.36	-0.07	0.28	0.36	○	16/17	8248	1244	6221	783	—	—
	R3	6075	.30	0.84	0.52	—	—	—	5743	0	451	5292	—	—
2R	R1	10583	.53	-0.64	0.49	-0.01	○	14/15	11073	10095	978	—	—	—
	R2	9417	.47	0.59	0.55	—	—	—	8927	488	8439	—	—	—

*OR1～5：規準設定前の所属ランク・グループ

表 15
RM-LR 法（GTM 分布指定なし）20,000 人リーディング・テスト・データ分析結果

		人数	比率	$\bar{x}$	SD	CP	可否	CRS	NSS	OR1	OR2	OR3	OR4	OR5
5R	R1	4280	.21	-0.94	0.34	-0.67	○	14/15	3957	3489	468	0	0	0
	R2	3824	.19	-0.43	0.19	-0.24	○	18/19	4373	778	2889	706	0	0
	R3	4091	.20	-0.03	0.18	0.17	○	21/22	3329	13	462	2519	335	0
	R4	3927	.20	0.36	0.18	0.64	○	26/27	4965	0	5	866	3471	623
	R5	3878	.19	1.01	0.40	—	—	—	3376	0	0	0	121	3255
4R	R1	5172	.26	-0.88	0.35	-0.56	○	15/16	4995	4459	536	0	0	—
	R2	4962	.25	-0.28	0.20	-0.04	○	20/21	5596	708	4090	798	0	—
	R3	5106	.26	0.21	0.19	0.52	○	25/26	5180	5	336	4135	704	—
	R4	4760	.24	0.92	0.41	—	—	—	4229	0	0	173	4056	—
3R	R1	6803	.34	-0.78	0.36	-0.37	○	17/18	7189	6385	804	0	—	—
	R2	7035	.35	-0.02	0.23	0.35	○	23/24	6663	418	5804	441	—	—
	R3	6162	.31	0.80	0.42	—	—	—	6148	0	427	5721	—	—
2R	R1	10458	.52	-0.57	0.42	0.00	○	20/21	10591	10063	528	—	—	—
	R2	9542	.48	0.57	0.47	—	—	—	9409	395	9014	—	—	—

MRM 分析や RM-LC 分析と異なり、RM-LR 法においては、ELP20,000 の L 部門、R 部門の2～5ランク分析のすべての隣接ランク・グループ間に適切な分割点をもたらす結果となった。

注目すべき特徴は、MRM や RM-LC 法に比べて、分割点が狭い得点域に集中していることであろう。例えば、表8（RM-LC 法 L 部門）の5クラス分析と表 14（RM-LR 法 L 部門）の5ランク分析の結果を比較すると、、RM-LR 法の分析では、「分割点の枠外」の R1 は 0～10 点、R5 が 19～20 点なのに対して、「分割点の枠内」の R2 は 11～12 点、R3 が 13～15 点、、R4 が 16～18 点と、枠外と枠内の得点域に大きな差が出ている。一方、RM-LC 法では、枠外の C1 が 0～8 点、C5 が 25～29 点、枠内の C2 が 9～14 点、C3 が 15～19 点、C4 が 20～24 点と、枠外と枠内の得点域に大きな差は見られない。

実際の規準設定の状況では、習熟度最上位層のグループ、最下位層のグループにできるだ

け多くの学習者を選抜することが目的の場合は、今回の RM-LR 法の結果が適用する可能性があるが、各グループ内の習熟度の差をできるだけ小さくすることが目的の場合は、RM-LC 法のほうが適しているかもしれない。

他のサンプルサイズについては、GTM が適用できないサイズもあるため、SOM を使用することとし、これまでに行ってきた VKS データ分析と比較するため、SOM 推定における初期設定の「一様分布」で分析を行った。L、R 部門ともに、ELP20,000、10,000、5,000、1,000、500、250、100 の2値型データについて、実施した2～5ランク分析において、すべての隣接クラス間に、適切な分割点を確認することができた。

一方、多値型データの分析では、異なる結果となった。表 16 は、L20,000_P の多値型データを GTM (分布指定なし) の条件で分析した際の、結果をまとめたものである。2、3ランク分析では、すべての隣接ランク・グループ間で適切な分割点を確認できたが、4、5ランク分析では、妥当な交点が得られないケースもあった。

表 16
RM-LR 法 (GTM 分布指定なし) 20,000 人リスニング・テスト(多値型データ)分析結果

		人数	比率	$\bar{x}$	SD	CP	可否	CRS	NSS	OR1	OR2	OR3	OR4	OR5
5R	R1	6360	.32	-1.41	0.58	×	×							
	R2	1969	.10	-0.97	0.33	-0.96	○							
	R3	3049	.15	-0.70	0.18	-0.52	○							
	R4	2995	.15	-0.35	0.12	-0.16	○							
	R5	5627	.28	0.28	0.50	—	—							
4R	R1	7381	.37	-1.37	0.56	-0.82	○							
	R2	2070	.10	-0.76	0.24	-0.80	×							
	R3	4490	.22	-0.50	0.23	-0.20	○							
	R4	6059	.30	0.24	0.51	—	—							
3R	R1	8083	.40	-1.33	0.56	-0.87	○	3/4	5723	5723	0	0	—	—
	R2	4218	.21	-0.63	0.21	-0.37	○	4/5	4655	1991	2604	60	—	—
	R3	7699	.38	0.12	0.51	—	—	—	9622	369	1614	7639	—	—
2R	R1	10122	.51	-1.21	0.56	-0.61	○	4/5	10378	9426	952	—	—	—
	R2	9878	.49	-0.03	0.54	—	—	—	9622	696	8926	—	—	—

多値型データの ELP20,000、10,000、5,000、1,000、500、250、100 についても、「SOM 一様分布」の分析を行ったところ、表 17 が示す通り、いくつかのクラス間において、適切な分割点を得ることができなかった。

GTM による分析は、今回扱ったサンプルの中では、ELP20,000, 10,000, 5,000 にのみ適用できたが、これらの相対的に大きなサンプルを SOM によって分析すると、いずれのサンプルにおいても、4ランク分析の R2/R3 に対して適切な交点が得られなかった。また、5ランク分析では、いずれのサンプルでも、ランクが高いはずの R3 の平均値が R2 のものよりも低くなり、RM-LR 法では通常は想定されない順番を並べ替える必要が生じた。

表 17
RM-LR 法 リスニング・テスト(多値型データ)の分析結果―サンプルサイズによる影響

	L20,000_P				L10,000_P				L5,000_P				L1,000_P			
	R1/2	R2/3	R3/4	R4/5	R1/2	R2/3	R3/4	R4/5	R1/2	R2/3	R3/4	R4/5	R1/2	R2/3	R3/4	R4/5
2R	○				○				○				○			
3R	○	○			○	○			○	○			○	○		
4R	○	×	○		○	×	○		○	×	○		○	○	○	
5R	○	○	○	○	○	×	×	○	○	×	○	○	○	×	○	○

	L500_P				L250_P				L100_P			
	R1/2	R2/3	R3/4	R4/5	R1/2	R2/3	R3/4	R4/5	R1/2	R2/3	R3/4	R4/5
2R	○				○				○			
3R	○	○			○	○			○	○		
4R	○	○	○		○	○	○		○	○	○	
5R	○	○	○	○	○	×	×	○	○	○	○	○

2値型データでは、サンプルサイズに関係なく、SOM 一様分布の分析が機能したが、多値型データでは、「大型サンプル」でランク数が多い分析に問題が出る傾向が見られ、「小型サンプル」でも適切な交点が得られないケースが見られた。RM-LR 分析が MRM や RM-LC 分析よりも安定した結果を導くことは間違いないようだが、今回の多値型データの分析については、時には機能しないこともあることを示している。

クラス・ランク分布の指定について

ここまで述べてきた MRM、RM-LC、RM-LR の GTM 及び SOM による分析は、いずれもそれぞれの分析手法の初期設定に基づいて行ったものである。その結果、MRM、RM-LC、RM-LR の GTM については分布の形状は指定されず、RM-LR の SOM については一様分布が指定された。分布の形状を選定することは、規準設定の方向性に大きな影響を及ぼすものであるが、今回の研究では、初期設定の状態で、どの程度安定した結果を残せるかをまず探ることとした。

ELP20,000 の「GTM 指定なし」による RM-LR 分析の結果、リスニングとリーディングの2値型データの2～5ランク分析において、すべての隣接するランク間に適切な交点を確認することができたが、多値型データについては機能しない場合もあった。一方、GTM が使用可能なサンプルサイズの ELP20,000、10,000、5,000 を「一様分布」で分析すると、2値型、多値型データともに、意図したすべての分割点を設定することができた。

WINMIRA 2001 でも事前にクラス分布の確率を設定できるため、ELP20,000、5,000、500 の3つのサンプルについて各クラスの分布確率を均等化し、分布を指定しない初期設定の場合と規準設定の効果を比較することとした。表 18、19 は、リスニング、リーディングの2値型データ、表 20 は、多値型データの各クラスの MRM による分割点設定の成否の結果をまとめたものである。

2値型データについては、均等分布の ELP20,000 においてリスニング2～5クラス分析、リーディング2～4クラス分析のすべての対象クラス間に分割点を確認できたが、クラス間の平均値の差が小さく、明確な分割点とは言えないものも含まれている。また、サンプルサイズが小さくなると分割点設定が困難になる状況は、同じように分布形状を指定した SOM による2値型データの RM-LR 分析結果とは大きく異なる。

表 18
MRM リスニング・テストの分析結果―分布指定による影響

	ELP20,000 (L)				ELP5,000 (L)				ELP500 (L)			
	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5
2C 初期	○				○				○			
2C 均等	○				○				○			
3C 初期	○	○			○	○			○	○		
3C 均等	○	○			○	○			○	○		
4C 初期	○	○	○		○	×	○		×	○	○	
4C 均等	○	○	○		○	×	○		×	○	○	
5C 初期	○	×	○	○	○	×	○	○	○	×	×	×
5C 均等	○	○	○	○	○	○	○	○	○	×	×	○

表 19
MRM リーディング・テストの分析結果―分布指定による影響

	ELP20,000 (R)				ELP5,000 (R)				ELP500 (R)			
	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5
2C 初期	○				○				○			
2C 均等	○				○				○			
3C 初期	○	×			○	×			○	×		
3C 均等	○	×			○	×			○	×		
4C 初期	○	×	○		×	×	×		×	○	×	
4C 均等	○	○	○		○	○	○		×	○	×	
5C 初期	○	×	×	○	○	×	×	×	×	×	×	×
5C 均等	○	×	○	○	○	×	○	○	○	×	×	○

一方、表 20 が示すように、多値型データについては、ELP5,000 の2～5クラスのすべての均等分布の分析において、適切な交点が得られないクラス・ペアが確認された。分布を指定することで変化は見られるものの、特定の分布を指定しない初期設定の状況と比べて、顕著な改善には至らなかった。

表 20
MRM リスニング・テスト(多値型データ)の分析結果―分布指定による影響

	L20,000_P				L5,000_P				L500_P			
	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5
2C 初期	○				×				○			
2C 均等	○				×				×			
3C 初期	×	○			○	×			○	×		
3C 均等	○	○			○	×			○	○		
4C 初期	○	○	○		×	○	×		×	×	○	
4C 均等	○	○	○		×	○	×		×	○	○	
5C 初期	×	×	○	×	×	×	○	×	×	○	なし	なし
5C 均等	×	×	○	×	×	×	○	×	×	×	×	×

RM-LC 法についても同じ3つのサンプルに対して各クラスの分布確率を均等化し、分布を指定しない初期設定の場合と規準設定の効果を比較した結果、リスニングの2値型データについては、均等分布のすべての対象クラスにおいて適切な交点が得られた。一方、リーディングの2値型データについては、表 21 の通り、分布の均等化により、全体的な改善は見られたものの、4、5クラス分析で分割点が得られないクラス・ペアが見られた。

表 21
RM-LC リーディング・テストの分析結果―分布指定による影響

	ELP20,000 (R)				ELP5,000 (R)				ELP500 (R)			
	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5
2C 初期	○				○				○			
2C 均等	○				○				○			
3C 初期	○	○			○	○			○	○		
3C 均等	○	○			○	○			○	○		
4C 初期	○	×	○		○	×	○		○	○	○	
4C 均等	○	○	○		○	○	○		○	○	○	
5C 初期	○	○	×	○	○	×	×	○	×	○	×	○
5C 均等	○	○	×	○	○	○	×	○	○	○	○	○

RM-LC 法の多値型データについては、表 22 ように、ELP500 の5クラス分析において、均等分布の方に若干の改善傾向は見られたが、全体的には大きな差異は見られなかった。

表 22
RM-LC リスニング・テスト(多値型データ)の分析結果―分布指定による影響

	L20,000_P				L5,000_P				L500_P			
	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5	C1/2	C2/3	C3/4	C4/5
2C 初期	○				○				○			
2C 均等	○				○				○			
3C 初期	○	○			○	○			○	○		
3C 均等	○	○			×	○			○	○		
4C 初期	○	×	○		○	×	○		○	×	×	
4C 均等	○	×	○		○	×	○		○	×	×	
5C 初期	×	○	×	○	○	○	×	○	×	×	×	×
5C 均等	×	○	×	○	○	×	○	○	○	×	×	○

潜在クラスの分布を均等化した RM-LC 法の分析については、同様の措置を施した MRM 分析よりも安定した分割点設定の結果となったが、RM-LR 法の一様分布の精度には至らなかったと言える。

MRM による規準設定

Jiao et al. (2011) に基づき、我々が議論してきた規準設定法では、隣接するクラスの平均値間に交わる確率密度曲線の地点を適切な分割点として認識してきた。今回の研究では 20,000 人の言語テストデータ提供を受けて、大規模なサンプルサイズの分析を行う機会を得たが、それでも適切な分割点としての交点が得られない場合があった。交点が得られた場合でも、隣接するクラスやランク・グループのの2つの平均値のどちらかに近似する値となることもあり、2つの確率密度曲線の重複部分が大きくなる場合も少なくなった。上記の条件で適切な交点を観測できなかった場合でも、規準設定を行うことはできないだろうか。

ELP20,000 (R 部門) の MRM 3クラス分析を例に考えてみることとする。図3が示すように、C1、C2 の間に適切な分割点を確認することができるが、C2、C3 については交点すら得られない状況である。上記の条件に従うならば、ELP20,000 (R 部門) の MRM 3クラス分析による規準設定は成立しないことになる。

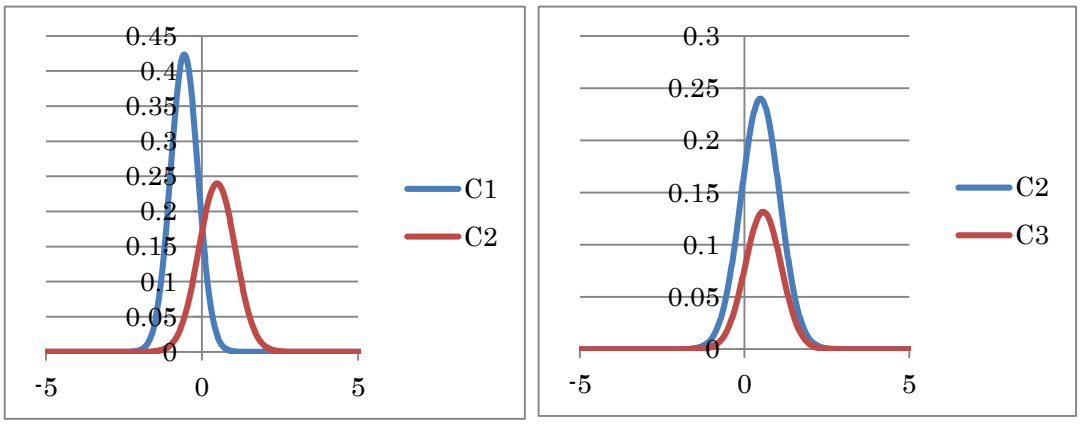

図3　ELP20,000 の R 部門 3 クラス分析

MRM や LCA の分析によって決定する潜在クラスは、受験者の能力パラメター以外の質的な要因に大きな影響を受けるとされる (Baghaei & Carstensen, 2013; Toker, 2016 等参照)。確率密度曲線の交点が得られた C1/C2、得られなかった C2/C3 の隣接する潜在クラス・ペアの平均正答率を3つの問題セクションに分けて計算すると、図4のような状況であることが確認できる。

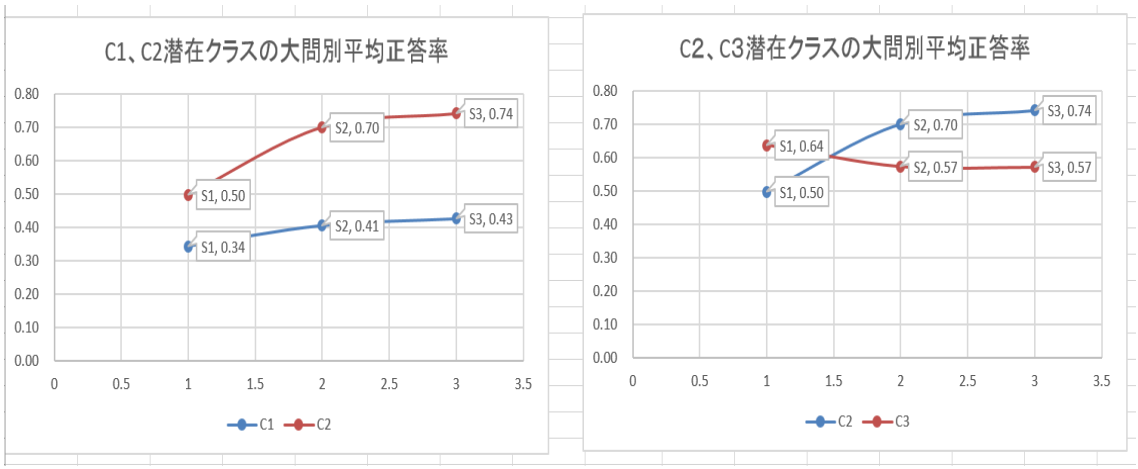

図4　ELP20,000 (R 部門) 3クラス分析の大問別平均正答率の対比

C1、C2 については、すべての部門で相対的に習熟度が高いと考えられる C2 が平均値において上回っているが、C2、C3 については、S1 (短文の語彙) 部門においては、全体の平均値が相対的に高い C3 が C2 よりも高い値を示しているものの、S2 (長文穴埋め)、S3 (長文内容理解) の問題においては、C2 の値が C3 を上回っている。

MRM 分析では、3つのクラスを設定する際に、単にテストの総合点だけでなく、テストの構成要素も加味して、クラス分割を行っている可能性があると考えることはできないだろうか。同様に、図5は、ELP20,000 のL部門の５クラス分析を、リスニングの3つの大問別に、5つのクラスに所属

する受験者の平均正答問数を比較したものである。問数が少ない S3 が、すべての潜在クラスで最も低い数値を示しているが、全体の平均値でわずかに C2 を上回る C3 （表2参照）が、S2 の問題正答数においては、C2 よりもやや低くなっていることがわかる。また、C2 までは問題数の同じ S1 と S2の正答数は均衡していたが、相対的に能力上位を占める潜在クラスほど、S2 の問題を S1 に比べて難しいと感じていたのではないだろうか。

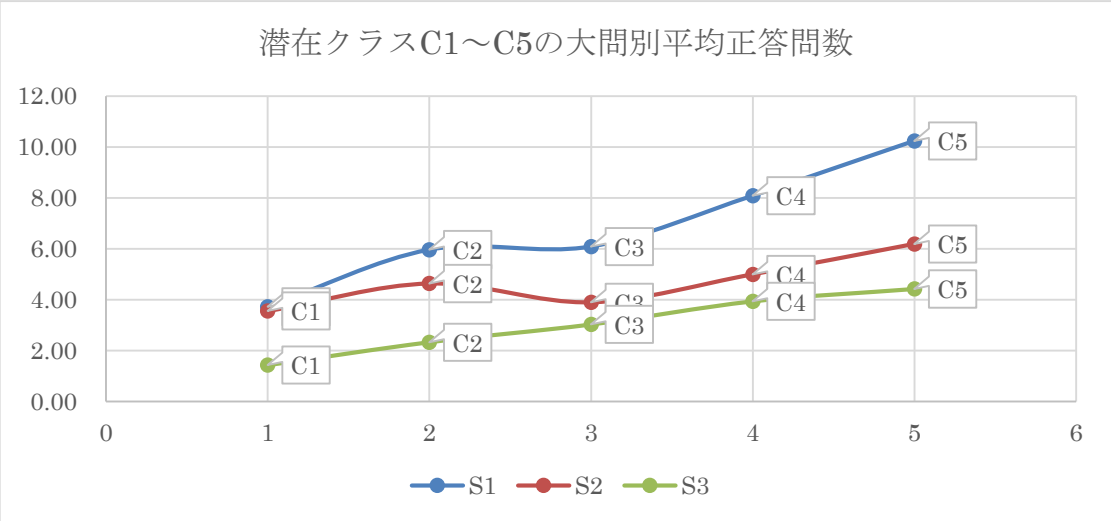

図5 ELP20,000 (L) 部門5クラス分析

このような潜在的な識別機能を内包している MRM を効果的に使用すれば、全体として同じような平均値を示す異なる集団に対して、語彙や長文読解等、グループの受講生のニーズに応じた指導を行うことも可能になるかもしれない。

最も安定していると考えられる RM-LR であっても、分布を指定した分析と指定しない分析とで、異なる結果が導かれたが、MRM についても、分布の指定の是非や意味について、検証を進めていく必要があるだろう。

MRM、RM-LC、RM-LR のどの方法を選ぶかを議論するよりも、それぞれの特徴を十分に理解して、それぞれの利点を規準設定やそのほかの目的に効果的に活用していく取り組みが今後も望まれる。

考察

研究課題1： MRM 分析における受験者サンプルサイズと分割点設定の効果

20,000 人の受験者で構成される英語能力試験のデータ （ELP20,000） の分析とともに、MRM 分析におけるサンプルサイズの影響を比較分析するために、様々な大きさの受験者集団を無作為に抽出し、リスニング、リーディングの技能別に全データの 20,000 人 (ELP20,000)、その一部の 10,000 人 (ELP10,000)、5,000 人 (ELP5,000)、1,000 人 (ELP1,000)、500 人 (ELP500)、

250 人 (ELP250)、100 人 (ELP100) のサンプルから形成されるデータを作成した。

これらのサンプルを比較すると、相対的に受験者数の大きなサンプルのほうが、安定した分割点設定が可能になる傾向は見られたものの、初期設定の状態で、ELP20,000 のリスニング(L)部門で2～4 クラス、リーディング(R)部門では2クラス分析のみ、対象クラスの規準設定を行うことができた。

法月 (2017) は、ELP20,000 よりもはるかに小さな 213 名の語彙テスト（VKS213）データを分析し、2、3クラス分析の分割点設定を行っているが、単純に受験者サンプルのサイズが大きくなれば、MRM による規準設定が容易になるとは言えないことが、明らかになった。

研究課題2　MRM による多値型データの分割点設定

20,000 人の受験者で構成される ELP20,000 のうち、L 部門の5つの説明文に対応するペア項目群への解答結果を、各ペア 2 問とも不正解の場合は 0 点、1 問正解の場合は 1 点、2 問とも正解の場合は 2 点と解釈し、多値型のテスト解答データ (L20,000_P) として分析を行った。

また、2値型データの分析の時と同じように、10,000 人、5,000 人、1,000 人、500 人、250 人、100 人のデータ（L10,000_P、L5,000_P、L1,000_P、L500_P、L250_P、L100_P）を抽出して、サンプルサイズの影響を比較し、分割点設定の効果を探った。

法月 (2017) が 193 名の語彙テスト理解度自己評価 (VKS193_P) の回答データに対して行った MRM 分析では、すべての2～5クラス分析のいずれの隣接クラス間でも適切な分割点が得られない状況であったが、L20,000_P で2、4クラス、L10,000_P と L250_P で2、3クラス分析が成立し、サンプルサイズにあまり影響されず、分割点設定の可否が生じる結果となった。一方で、L500_P、L250_P、L100_P の 4 クラス、5クラス分析で、C4 や C5 の構成員が 0 人になるなど、小さなサンプルになると、2値型データ以上に大きく分布が偏る傾向が見られた。

研究課題3　RM-LC 法による分割点設定

単純ラッシュモデル (RM)の分析と潜在クラス分析 (LCA) を個別に行い、確率密度曲線を通じて分割点設定を行う方法 (RM-LC 法) が実践可能か否かについて、L、R 部門別に、ELP20,000、10,000、5,000、1,000、500、250、100 データの2～5クラス分析を行った。

L、R 部門ともに、MRM よりも RM-LC 法において、全体的に分割点設定が成立するケースが多くなる結果となったが、特に L 部門で、ELP20,000 で2～5クラスすべての隣接クラスの分割点を設定することが可能だったのに対して、ELP10,000、5,000 では2～4クラス、ELP1,000、500、250 では2、3クラス、ELP100 では2クラス分析のみ成立する結果となり、サンプルサイズの影響が顕著に表れた。

一方、R 部門では ELP1,000、500 が2～4 クラス、その他の5つのサンプルサイズにおいてはいずれも2、3クラスのみ分割点設定が成立する結果となり、サンプルサイズの影響はあまり感じられない状況が示された。

ちなみに、RM-LC の分析では、RM-LR 同様に、同じ素点が別のクラスに分類されることはあるが、今回分析に使用した RM の値と素点は、常に一対一の関係にある。これに対して、MRM の分析では、同じ素点の受験者であっても、クラスだけでなく、受験者パラメター値が異なる場合もある。例えば、ELP500 の R 部門で、41 問中 18 問正解の受験者は、RM-LC 法の5クラス分析では、C1、C2、C3、C4 の4つのクラスに分かれているが、受験者パラメター値は常に-0.261 である。同じ正解数の受験者の MRM 分析の結果を見ると、C1、C2、C3 の3つのクラスに分かれ、受験者パラメター値も、-0.17、0.03、0.06 と変動している。

Templin & Jiao (2012) は規準設定の観点から LCA 分析について議論しており、RM-LC 法の可能性を探求する意義も大いにあると考えられるが、そのためにも、Toker (2016) のように、LCA と MRM のクラス分類における特徴の相違点について探求していく必要があるだろう。

研究課題4: RM-LR 法による分割点設定

単純ラッシュモデル(RM) と潜在ランク理論 (LRT) の分析を個別に行った後に、確率密度曲線を計算して、分割点設定を行う RM-LR 法についても、L 部門と R 部門の2値型データ、L 部門の多値型データのそれぞれ7つのサンプルの分析に適用することとした。

分析の結果、2 値型データについては、L 部門と R 部門ともに、すべてのサンプルの2～5ランク分析（SOM 一様分布）において、有効な分割点設定を行うことができたが、多値型データについては、いくつかの隣接ランク間で適切な分割点設定ができない状況が確認された。また、2値型データについても、ELP20,000 について行った「GTM 指定なし」のモードによる多数クラスがかかわる分析では、「分割点枠内」の中間層の得点域が狭く、「分割点枠外」の上位層や下位層の得点域が広くなる傾向が確認されており、規準設定の目的によっては、意図した習熟度区分が効果的に機能しない可能性が考えられる。

大型サンプルでは GTM による分析が可能だが、ELP20,000、10,000、5,000 のサンプルに対して行った GTM の初期設定の「（分布の）指定なし」の分析では、多値型データの問題点は解消されなかった。その一方で、GTM についてもランク・グループの分布が均等化される「一様分布」を選択すると、多値型データでもすべての分割点設定が有効に機能することがわかった。

GTM による分析は、今回使用したサンプルとしては、1,000 名以下のものには適用できなかったため、LP1,000_P、L250_P の5ランク分析の規準設定の解決にはつながらなかったが、確率密度曲線を使った規準設定において、RM-LR が、RM-LC や MRM の分析よりも有効性が高かったことは間違いないだろう。

今回 LRT 分析に使用した Exametrika には、「一様分布」、「指定なし」以外に、「正規分布」の分布の形状を選択することができるが、それぞれの分布状況と規準設定の関係についても、追究していく価値がある。

クラス・ランク分布の指定について

RM-LR 法の分析では、各ランク・グループの分布は「指定なし」よりも均等化される「一様分布」のモードを選択することで、多値型データの分割点設定に適用できる可能性が示唆された。MRM や RM-LC 法でも、分布の確率を均等化する設定を 20,000、5,000、500 人の2値型、多値型データについて行ったところ、特に2値型データの分析では顕著な改善が見られたが、多値型データについては大きな効果は見られなかった。

いずれにしても、実践的な規準設定においては、何の目的で規準設定を行うかを十分に考慮したうえで、分布を指定すべきか否か、指定する場合はどのような分布を指定するか等の問題に加えて、潜在クラス数の選定についても、合理的に検証することが求められる。10,000 人のシミュレーションデータを使った Jiao et al. (2011) は、各潜在クラスの所属者数を能力の低い方から、600、1,300、5,300、2,200、600 人と設定して、各クラスの閾値等を指定して、規準設定の分析を行った。今回の分析では技術的にこのような細かな分布指定を行うことはできなかったが、今後の分析において追究する価値がある。

Jiao et al. (2011) は、分割点の数値として true scores と estimated scores を併記しているが、今回の研究では、分析の便宜上、後者の数値を軸に議論を進めてきた。今後の研究においては、特に分布を指定する場合、後者の方法では分布に誤差が生じるため、前者のデータをベースにした分析結果と比較することも検討すべきであろう。

MRM は規準設定に適さないのか？

確率密度曲線を使った MRM による規準設定は、RM-LR 法や RM-LC 法に基づく方法に比べて、分割点の設定のハードルが高いのではないだろうか。法月（2017）が分析した VKS213 は、英語の単語に関する知識を問う単一問題形式のテストであったが、今回の研究で分析した ELP20,000 は、L、R 部門ともに複合形式の総合的な英語能力のテストであった。単語力はあっても読解力は劣っていることもあれば、その反対の状況の学習者もいる。モノローグ形式の説明とダイアログ形式の会話の聴き取りに対して、大きな心理的な負荷の違いを感じる学習者もいる。単純に合計スコアは均衡していても、習熟度状況や学習背景等の要因が解答様式に大きな差異を及ぼすこともある。

MRM がこのような質的な特徴の差異を見極めるうえで、他の統計的手法にない「敏感さ」を持っているとするならば、一見して安定感に欠ける「変動性」こそが、MRM の規準設定における統計的手法の可能性を示唆するものとなるかもしれない。

唯一の規準設定の手段として MRM を位置づけることには無理はあるかもしれないが、RM-LR、RM-LC 等とともに研究を重ね、より洗練された規準設定の統計的解決モデルを構築していくことが望まれる。

おわりに

この研究の目的は、「分割点の設定」であり、「分割点は客観的に決めることはできないが、客観的に適用することができるもの」であることを追求している。ラッシュモデル (RM)、潜在ランク理論 (LRT)、潜在クラス分析 (LCA)、混合ラッシュモデル (MRM) のうち、どの手法が適切なのか、あるいは、組み合わせによる手法や、まったく別の方法が考えられるのか。分割点を科学的で、明確かつ唯一の統計量として計算し、設定することはできるのだろうか。新井 (2014) は、中世以降の新しい分析法として「数学」が発展してきた背景を述べ、イギリスの数学者・論理学者のチューリングという人の「計算できる」とはどういうことかの定義を紹介している。それは、「有限の知識、特定の条件の下における特定の手続き、同様に繰り返す」というたった 3 つの要素であるという。そして、数学の一分野である確率と統計について「正しくはないが、意外と当たる」と説明している。

我々の研究が追求する分割点設定の「計算」においても、有限の知識として様々な統計手法を検討し、特定の条件下で得られた複数のデータ・セットを使用し、特定の統計手法の比較検討を、同様に繰り返すことで、「意外と当たる」確率をより高めていく必要がある。

参考文献

Baghaei, P. & Carstensen, C. H. (2013). Fitting the mixed Rasch model to a reading comprehension test: Identifying reader types. *Practical Assessment, Research & Evaluation. 18* (5). Retrieved from http: //pareonline.net/getvn.asp?v =18&n=5.

Bramley, T. (2010). *Locating objects on a latent trait using Rasch analysis of experts' judgments.* A paper presented at the conference "Probabilistic Models for Measurement in Education, Psychology, Social Science and Health," Copenhagen, Denmark (June, 2010). Retrieved from http://www.cambridgeassessment.org.uk/ca/digitalAssets/186233_TB_locating_objects_Rasch2010.pdf

Cizek, G. J. (2006). Standard Setting (p. 226). in Downing, S. M. & Haladyna, T. M. (Eds.) *Handbook of Test Development.* Lawrence Erlbaum Associates, Publishers.

Cohen, A.S., Wollack, J.A., Bolt, D.M., Mroch, A.A. (2002). 'A mixture Rasch model analysis of test speededness'. A paper presented at the annual meeting of the American Education Research Association, New Orleans, LA. Retrieved from https://testing.wisc.edu/research%20papers/AERA%202002%20(Cohen,%20Wollack,%20&%20Mroch)

Jiao, H., Lissitz, B., Macready, G., Wang, S., & Liang, S. (2011). Exploring using the mixture Rasch model for standard setting. *Psychological Test and Assessment Modeling, 53,* 499-522.

Kelderman, H. & Marcready, G.B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. *Journal of Educational Measurement,* 27, 307-327.

Kreiner, S., Hansen, M., & Hansen, C.R. (2006). On local homogeneity and stochastically ordered mixed Rasch models. *Applied psychological measurement 30,* 271-297.

Kreiner, S. (2007). Determination of diagnostic cut-points using stochastically ordered mixed Rasch models. In von Davier, M., & Carstensen, C.H., (Eds.). *Multivariate and mixture distribution Rasch models: Extensions and applications.* (pp.131-146). New York: NY: Springer.

Lee, Y-H., & Chen, H. (2011). 'A review of response-time analyses in educational testing'. *Psychological Test and Assessment Modeling, 53.* (pp.359-379).

Mislevy, R.J., & Verhelst (1990). Modeling item responses when different subjects employ different solution strategies, *Psychometrika,* 55(2), 195-215.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement,* 14, 271-282.

Rost, J., & Langeheine, R. (1997). 'A guide through latent structure models for categorical data'. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp.13-37). Munster, Germany: Waxmann.

Shojima (2007). Neural test theory. DNC Research Note, 07-02. Retrieved from http://www.rd.dnc.ac.jp/~shojima/ntt/index.htm

Templin, J., & Jiao, H. (2012). Applying model-based approaches to identify performance categories. In G.J. Cizek (Ed.), *Setting performance standards. (Second Edition) (pp.379-397). New York, NY: Routledge.*

Toker, T. (2016). A comparison of latent analysis and the mixture Rasch model: A cross-cultural comparison of 8th grade mathematics achievement in the fourth international mathematics and science study. (Doctoral dissertation). Retrieved from https://digitalcommons.du.edu/cgi/viewcontent.cgi?article=2172&context=etd

von Davier, M. (2001). WINMIRA [Computer software]. Groningen, the Netherlands; ASC Assessment Systems Cooperation. USA and Science Plus Group.

Zieky, M.J. , Perie, M., & Livingston, S.A. (2008). *Cutscore: A Manual for Setting Standards of Performance on Educational and Occupational Tests, Princeton, NJ: Educational Testing Service.*

新井紀子. (2014). 『ロボットは東大に入れるか』. イースト・プレス.

植野真臣・荘島宏二郎. (2010). 『学習評価の新潮流』. 朝倉書店.

大友賢二 (研究代表; メンバー: 伊東祐郎・法月健・藤田智子・渡部良典). (2012). 『言語テストの規準設定 報告書』. 2011 年度 公益財団法人英語検定協会英語教育センター委託研究.

大友賢二 (研究代表; メンバー: 伊東祐郎・法月健・藤田智子・渡部良典). (2013). 『言語テストの規準設定 報告書 第 2 号』. 2012 年度 公益財団法人英語検定協会英語教育センター委託研究.

大友賢二 (研究代表; メンバー: 伊東祐郎・法月健・藤田智子・渡部良典). (2014). 『言語テストの規準設定 報告書 第 3 号』. 2013 年度公益財団法人英語教育センター委託研究.

大友賢二 (研究代表;メンバー: 池田 央・村木栄治・中村洋一・法月 健). (2015). 『ICT 等を活用した評価についての調査・研究報告書』. 2014 年度公益財団法人英語教育センター委託研究.

大友賢二・中村洋一・法月健. (2016). 『Mixture Rasch Model による英語能力の規準設定 報告書』. 2015 年度公益財団法人英語教育センター委託研究.

大友賢二. (2016). 『Mixture Rasch Model による英語能力の規準設定 報告 (2016 年 3 月 18 日特別講演会、2016 年 9 月 動画配信). 2016 年度公益財団法人英語教育センター委託研究.

大友賢二・中村洋一・秋山實. (2008). Test Data Analysis Program Ver. 2.02.

小泉利恵・飯村英樹. (2010).「ニューラルテスト理論の特徴: 古典的テスト理論・ラッシュモデリングとの比較から」. 『日本言語テスト学会紀要』13 号 (pp. 91-109).

荘島宏二郎 (2011). Exametrika (Version 5.3) [Software] Available from http://antlers.rd.dnc.ac.jp/~shojima/exmk/jindex.htm

荘島宏二郎. (n.d.).「ニューラルテスト理論」. Retrieved from http://www.rd.dnc.ac.jp/~shojima/ntt/jindex.htm

法月健. (2012a).「課題規準設定におけるラッシュモデルの有用性」. in 大友賢二 (研究代表). (2012). 『言語テストの規準設定 報告書』. 2011 年度 公益財団法人英語検定協会英語教育センター委託研究. (pp. 117 – 126).

法月健. (2012b). 「規準設定における潜在ランク理論の有用性 — 項目応答理論と古典的テスト理論との比較」. in 大友賢二 (研究代表). (2012). 『言語テストの規準設定 報告書』. 2011 年度 公益財団法人英語検定協会英語教育センター委託研究. (pp. 117 – 126)

法月健. (2013). 「受容語彙を測定するプレイスメントテストにおけるラッシュモデルと潜在ランク理論に基づく規準設定の試行」. in 大友賢二 (研究代表). (2013). 『言語テストの規準設定 報告書 第 2 号』. 2012 年度 公益財団法人英語検定協会英語教育センター委託研究. (pp. 81 – 103).

法月健. (2014). 「実用英語検定の級別頻出単語に基づく英語受容語彙力テストの開発と規準設定」. in 大友賢二 (研究代表). (2014). 『言語テストの規準設定 報告書 第 3 号』. 2013 年度公益財団法人英語教育センター委託研究. (pp. 77 – 101).

法月健. (2017).「Mixture Rasch Model による英語能力の規準設定: 検討結果と今後の課題」 未出版原稿.

三輪哲. (2009). 「計量社会学ワンステップアップ講座 (3) 潜在クラスモデル入門」 in『理論と方法 (Sociological Theory and Method) 』 Vol. 24, No. 2: 345-356, Retrieved from https://www.jstage.jst.go.jp/article/ojjams/24/2/24_2_345/_article/-char/ja/.