公益財団法人　日本英語検定協会

２０１５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
報告書

２０１６年３月３１日

大友賢二 (筑波大学名誉教授)
中村洋一（清泉女学院短期大学教授）
法月　健（静岡産業大学教授）

公益財団法人　日本英語検定協会
２０１５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

報告書

１．　中間報告：大友賢二・中村洋一・法月　健

１．１．規準設定における統計処理
１．２．混合ラッシュ・モデル
１．３．**MRM** を使った規準設定手順
１．４．**MRM** 分析の今後の課題

２．　進捗状況報告：大友賢二・中村洋一・法月　健

２．１．規準設定に関する方法と統計処理
２．２．**MRM** の利点
２．３．**MRM** を使った規準設定手順
２．４．**MRM** 分析の今後の課題

３．　検討結果と今後の課題：

３．１．受容語彙能力テストの分析と今後の課題：法月　健
３．２．検討必要性の再確認：中村洋一
３．３．検討結果と今後の課題：大友賢二

Mixture Rasch Model による英語能力の規準設定
２０１５年度：中間報告

大友賢二・中村洋一・法月健

1. 1. 規準設定における統計処理:
Latent Class Analysis (LCA) と Mixture Rasch Model (MRM)

本研究は、CEFR の “3 Common Reference Levels” で “a common framework scale should be objectively determined in that they are based on a theory of measurement (p. 21) と指摘されている standard setting の統計的な方法論に関するものである。「目標の内容、その設定方法を検討する際には、それと同時に、目標の到達と未到達はどうすれば判断できるのかという課題に触れる必要」があり、「それができなければ、単なる表面的な解決に終わってしまう」 (大友監修, 2009, p. 2) との懸念に対応していくのが、我が国の英語教育における急務であろう。

CEFR の Can Do の指標検討では、主に “Item response theory (IRT) or ‘latent trait’ analysis” が使用された (pp. 207 - 212) とある。また、Dr. Matthias von Davier (MVonDavier@ETS.ORG, Senior Research Director, Educational Testing Service) は、2014 年 1 月台北で行われたワーショップの ppt 資料の中で以下のように指摘し、

- Classical standard setting methods do so by cut-scores (defined by experts) on a latent or observed test score scale
- Model-based levels and standards may be developed using statistical and psychometric models with ordinal latent variables

Model-based methods として、次のモデルを解説している。

Latent Class Models

Mastery Models, Mokken, DINA, ...

Ordered Latent Class / Discrete IRT Models

Mixture IRT and General Diagnostic Model

Model-Based Levels and Standards

Standard setting の状況に応じて、適用すべきモデルの選択肢はいくつかあるようであるが、本研究では、規準設定における統計処理のひとつとして Mixture Rasch Model の可能性をまず検討していきたい。次項では、関連する Rasch Model (RM)、 Latent Class Analysis (LCA)、 Mixture Rasch Model (MRM) の 3 つの理論について、順番に概観して行く。

1.1-1. ラッシュ・モデル (Rasch Model：RM)

ラッシュ・モデル (Rasch Model: RM) は、項目応答理論 (item response theory : IRT) の一つと考えられている。テストデータの処理できわめて重要なことは、大友 （1996, pp. 17-20） で述べているように、(1) 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)、(2) 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)、そして、(3) 能力ごとにわかる測定の精度 (multiple reliability estimation)である。Hambleton, Swaminathan, & Rogers (1991, p. 5) で示されている以下の 5 項目を満たす “alternative test theory”、つまり IRT が必要である。

(a) item characteristics that are not group-dependent, (b) scores describing examinee proficiency that are not test-dependent, (c) a model that is expressed at the item level rather than at the test level, (d) a model that does not require strictly parallel tests for assessing reliability, and (e) a model that provides a measure of precision for each ability score.

このようなことは、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ものである。このように、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。IRT においては、項目特性曲 (item characteristic curve: ICC) を用いることで、これを可能にした。ごく簡単に言えば、RM が必要なことは、以上のようなことが可能である理論であるからである。

デンマークの数学者である Rasch, G. によって提案されたこのモデルは、ラッシュ・モデル(Rasch model) と呼ばれ、単純ラッシュ・モデルとも言われ、その数式は $P = \exp(\theta - b) / (1 + \exp(\theta - b))$ <数式１> で示される。能力を θ、項目困難度を b として、能力が θ である受験者が困難度が b である項目に正答する確率を P とすると、この数式が当てはまるというものである。

項目応答理論は、Lord (1952) でその基礎が確立され、Lord & Novick (1968) で数理的に体系化されたものである。これは、TOEFL など、米国のテスト研究機関である ETS (Educational Testing Service) で開発実施されるテストを中心にすでに用いられてきていたものであり、米国のみならず、世界のいたるところで、言語テストを支えるテスト理論として、広く用いられてきているものである。この項目応答理論の主だったモデルとしては、One parameter logistic model： 1PLM, Two parameter logistic mode: 2PLM, Three parameter logistic mode: 3PLM が用いられているが、1PLM の Item Characteristic Curve: ICC は、$P = 1 / (1+\exp(-(\theta - b)))$ <数式２>と示されることがある。ラッシュ・モデルは、そのモデルの発想が項目答理論とは全く異なるものである。しかし、このモデルは、<数式１> の分母

分子を exp (θ－b)で割ると、次のようになり、<数式２> と一致する。そして数理モデルとして、ラッシュ・モデルは、項目応答理論の１PLM と一致することになる。

$$P = \exp(\theta - b) / (1 + \exp(\theta - b))$$
$$= \exp(\theta - b) / \exp(\theta - b) / ((1 + \exp(\theta - b)) / \exp(\theta - b))$$
$$= 1 / ((1/\exp(\theta - b)) + 1)$$
$$= 1 / (\exp (-(\theta - b)) + 1)$$
$$= 1 / (1 + \exp(-(\theta - b)))$$

したがって、項目応答理論の中でも、1PLM というよりもラッシュ・モデルと呼ばれることが多いので、ここでも Rasch Model という表現を用いることとする。

1.1-2. 潜在クラス分析 （latent class analysis: LCA）

我が国の文献では、「潜在クラス分析」という用語を見つけるのは、きわめて困難であるが、岡本・中獄・村瀬・山本 編 （1971, p. 167） では、「潜在構造分析 （latent structure analysis: LSA）」の中で、以下のような説明が見られる。

> 潜在構造分析方は 1950 年ラザースフェルド （P.F. Lazarsfelt） によって示された。基本的な考えは次のようである。測定しようとする事柄について○×式質問項目が使用されたとする。そのとき個々の質問に対する正答率、あるいは賛成率がデータとして得られるが、それは必ずしも調査者の知りたいと目指すものではない。本当に知りたいのは、一連の反応の背後にある潜在的な態度の構造なのである。
>
> 図１の実線の分布のことを「潜在的連続体」というが、このような連続分布の形では、解法が得られていないので、その程度に応じていくつかのクラス （これを潜在クラスという） に分ける。たとえば、好意群・非好意群の２群に分けることによって解いていく「潜在クラス分析 （latent class analysis: LCA）」が一般によく使われる。

この文献のほか、依田 監修 （1977, pp. 505-506） における「潜在構造分析 （latent structure analysis）」、東・梅本・芝・梶田 編 （1988, pp. 369-3） における「潜在構造分析 （latent structure analysis）」では以下のように示されている。

> 潜在クラス分析では、被検査者群が、潜在的で直接観察できない層（潜在クラスという）に
>
> 層別されていて、各層ごとに、調査や監査の項目への応答率が異なるものと仮定する。被験
>
> 者全体としての各項目への反応のデータを用いて、潜在クラスの個数、潜在クラスの相対的

な大きさ、潜在クラス別の各項目への応答率を推定しようとする手法である。

Templin & Jiao (p. 384) は、LCA を以下のように定義・説明している。

Latent class analysis (LCA) is a statistical procedure used to model unobservable (latent) groups believed to be underlying observed data. As described by Dayton and Macready (2007), LCA is closely related to the discrete mixture model and classical factor analysis. The simplest forms of LCA is that used for categorical item responses.

Both IRT and LCA are latent variable models. However, unlike IRT models where a continuous latent trait underlies a person's item performance, LCA uses latent discrete class membership to define a person's responses. In LCA, item response probabilities could be different across classes but item responses are independent within classes.

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との指摘は、興味深いものである。また、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いう指摘は、LCA を理解する上で重要なポイントである。

さらに、LCA の利点としてあげている次の指摘も重要である (pp. 383-384)。

The main benefit of LCA or extended model-based approaches to standard setting is to reduce the classification error associated with the subjectivity built into the current common practice of standard setting by making use of the information in item response patterns along with additional standard setting information. Although we began our chapter discussing ways in which classification models can (and should) incorporate human judgment from the standard setting process, by nature, classification models do not need such judgment.

上記 Templin & Jiao と同様の概念を三輪 (p. 345) は、以下のように表現している。

この半世紀ほどの統計データ解析の発展の中でも、潜在構造の分析法、あるいは潜在変数の利用は大きな発展をみた分野の 1 つである。潜在変数を用いることの利点は様々あるが、それらのうち主要なものとしては、節約の原理に適うことと、希薄化の修正ができることが挙げられる。前者は、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を、また後者は、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それぞ

れ意味する。

(中略)潜在クラスモデルとは、カテゴリカルな観察変数の背後にカテゴリカル潜在変数があ

ることを仮定して潜在構造を読み解くモデルをいう。

潜在クラス分析 （LCA） は、潜在クラスを仮定することによって、観測できるデータに潜んでいるが、しかし観測されない、つまり潜在する本質の究明を目指し、観測値を解釈しやすく、かつ高精度で「縮約」する統計的手法の枠組みであり、standard setting の Cut Score を設定する作業において大変有益な手法であると言えるであろう。

1.1-3. 混合ラッシュモデル (Mixture Rasch Model: MRM)

通常のラッシュモデルの分析では、「単一の構成概念が階層的な連続体を形成する項目の基盤となる」（Bond & Fox, 2007, p. 314）ことを仮定する一元性 (unidimensionality) が分析の前提条件となる。しかしながら、テストの構成や内容によっては、特定の受験者に不利な影響を与えたり、それらの項目難易度が正確に推定できなくなる場合もあり、社会背景や学習環境によって特定の受験者グループに有利あるいは不利に働く項目が存在する場合は、一元性を仮定した分析が効果的に機能しないこともある。

このような心理測定的な制約に対応するため、Mixture Rasch Model (MRM：混合ラッシュモデル) を追究した様々な研究が行われてきた (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)。

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。

本報告では、実際に MRM を使って試行分析を行い、その規準設定における可能性について、過去の研究で行われた他の統計モデルを活用した統計的解決法の結果と比較しながら、議論を進める。

1. 2. 混合ラッシュ・モデル （Mixture Rasch Model: MRM） 研究の展望

わが国における MRM の研究は、極めて少ないが、注目すべき事項も少なくない。その中では、植野・荘島 （2010） の述べている「潜在ランク理論 （latent rank theory: LRT）」、また、その応用を試みた法月 （2013）、（2014） などがある。ただ、外国における研究

で論じられている LCA とわが国で述べられている LRT の比較検討、それが、規準設定のための MRM とどう結びつくのかなどの検討は必要であろう。さらに、MRM のコンピュータ・ソフトである WINMIRA 2001 (von Davier, M. 2001) などの適応性は検討に値する。そうしたソフトを用いて、目的とする規準設定が、十分可能であるかの検討も重要な課題の一つである。

1.2-1. MRM の推定方法と処理プログラム

Templin & Jiao (2012, p. 388) は、estimation methods for the MRM として

① the marginal maximum likelihood estimation (MMLE) method with the expectation-maximization (E-M) algorithm (used in mdltm)"

② M-Plus

③ the conditional maximum likelihood estimation method (used in Winmira)

④ the Markov Chain Monte Carlo estimation method (used in WINBUG)

をリストアップしている。本研究で取り扱うテストデータの推定方法として、どの方法が適切なのか、それぞれの推定法の評価を調査していく必要があろう。

さらに、それぞれ特徴のある推定方法を採用しているプログラムのうち、どれが有用なのかについても検討が必要である。Templin & Jiao (2012) は the Mulitdimensional Discrete Latent Trait Model (mdltm) software を使用して levels of performance の研究をしている (p. 388)。Baghaei & Carstensen (2013) は、WINMIRA を使用して a reading comprehension test composed of 20 multiple-choice items の分析を行い、"Item fit for each class" といった観点からの分析も行っている。WINMIRA は Rost (1996, p. 459) のような比較的初期の文献にも紹介されており、また、ウェッブページでその内容に関する情報や、マニュアルが入手可能で、Kagi on line store から購入も可能であり、検討のとりかかりとしては、魅力的である。いくつかのモデルを比較したシンポジウムの overview (Rupp, 2009) や Rasch Measurement Analysis Software Directory: http://www.rasch.org/software.htm といったウェッブページもあり、プログラムについては、そのような情報を糸口に検討を深めていきたい。

1.2-2. WINMIRA 2001

WINMIRA は、Davier が開発を手がけたもので "Windows-program MIxed RAsch model (1995)" (Rost, p. 457) と紹介されている。また、WINMIRA のホームページでも、WINMIRA 2001 is a software for estimating Rasch Models, Mixture Rasch Models, Latent Class Models and Hybrid models. It provides models for dichotomous and polytomous data and can handle ASCII (text

file) and SPSS (sav) sources. It is available for windows operating systems. (http://www.von-davier.com, retrieved in Sep. 14, 2015) と掲載されている。

WINMIRA の情報は http://208.76.84.140/~svfklumu/wmira/index.html に掲載されており、またマニュアルも入手可能である (http://208.76.84.140/~svfklumu/wmira/winmiramanual.pdf)。

マニュアルには以下の紹介があり、幅広い用途が期待できる。

The Mixed−Rasch Model extends the Rasch model to a discrete mixture model. The main goal of applying this model is to classify a possible inhomogeneous sample into Rasch−homogenous subsamples.

The Mixed Rasch model can be used for very different tasks, e. g.

- for testing model fit of the Rasch Model (by comparing the one−class and the two−class solution),
- for identifying a Rasch scaleable subpopulation (or separating a class of unscaleables, respectively),
- for analyzing rating data, when different subsamples have different response sets,
- for measuring a latent ability, when different people apply different solution strategies for solving the items, or
- for profile analysis of questionnaire items with ordinal response formats.

パラメータの推定法はいくつか提案されているが、“The parameter in the mixed Rasch model and its generalizations to ordinal data can be estimated by means of an extended EM-algorithm with conditional maximum likelihood estimation of the item parameters in the M-step (Rost, p. 455).” とあり、WIMIRA でも “The latent classes are identified by means of an EM−Algorithm and the item − or threshold parameters are computed by means of conditional maximum Likelihood (CML) estimation within each M−step. The CML estimation requires the latent score distributionslatentscore, i.e., the distributions of test scores in each latent class, to be estimated in order ‘to condition out’ the person parameters in the CML−procedure.” とマニュアルにあるように、Rost と同様に、評価の高い推定法を採用している。

また、開発者の Davier にメールで問い合わせたところ、“I successfully tested WINMIRA 2001 with windows 10.” との返信があり、Windows 10 での稼働も可能であることが分かった。

本研究では、Windows 10 の環境を揃え、KAGI のウェッブページから(http://order.kagi.com/cgi-bin/store.cgi?storeID=1HN&&) 購入できる WINMIRA 2001 (academic single user license; $200, または academic PC Pool license (up to 10 installs); $800) を用いてデータの分析をしていきたい。

1. 3. MRM を使った規準設定手順

MRM に関する研究は数多く行われているが、MRM を使った規準設定の研究は、それほど多いとは言えない。そこで、Jiao, et al. (2011) の研究で行った分割点設定の算出方法に基づき、実際の言語テストデータを分析して、実践可能な規準設定手順を探ることとした。

1.3-1. 法月 (2014) との比較

大友賢二研究代表の下にまとめられた「言語テストの規準設定報告書 3」中の法月 (2014) の報告で使用した英検の級別に対応した 2 種類の受容語彙力テスト 50 問・自己評価 50 問の評価システム (VKS1 & VKS2) のうち、4 級～準 1 級に対応した VKS1 のテスト部門 50 問の 213 名のデータ（以降、VKS データ）を WINMIRA 2001 を使用して、2～5 クラスの MRM 分析を行い、法月 (2014) で採択したラッシュモデルと潜在ランク理論を併用した規準設定法（以降、RM-LR 法）の結果と比較することとした。

MRM 分析を正確に実施するには 1 万人規模のデータが必要であることを示唆する論文 (Jiao, et al. 2011) もあり、今回のような数百名以下のデータ分析では大きな測定誤差を生じ、十分に適応しない状況も想定された。対応策として、実データを基にシミュレーションデータを生成することで、大規模データの有する必要条件を満たす分析を行うことが可能と思われる Mplus (Muthen & Muthen, 2012) や mRm ((Preinerstofer & Forman, 2012) の様なソフトウェアの使用が考えられるが、相応のプログラミング技術を有することが効果的に実践するための最低条件と言える。結局、データが小さくても Bootstrap 機能でデータの適合度を調べることができ、SPSS のインターフェイスでデータ処理を行うことができる WINMIRA 2001 (von Davier, 2001) を今回の研究では使用することとした。

2～5 クラスの MRM の分析を行うと、いずれの場合も、サンプル数が小さいため通常の AIC や BIC 等の適合度指数は参照せず、Bootstrap 分析を行うことを指示する同一の警告が標示された。Bootstrap 分析を行うと、使用が勧められる 2 つの統計指標 (empirical p-values of the Pearson X^2 と Cressie Read Statistics) はいずれも 0.000 の値を示した。

表 1 は、法月 (2014) で使用した VKS データを使って、5 クラスの MRM 分析から得られた基本統計値を報告したものである。WINMIRA 2001 では、割り当てられた受験者の所属数が最も多い Class 1 から降順にクラス名が設定されるが、他の研究で一般的な結果提示に倣って、能力の平均値が低い方から昇順にクラス名を再配列した。法月 (2014) の潜在ランク理論のランクとの相関 (Spearman's rank correlation) を計算すると、 .853 を示した。

表１　5 クラスの MRM 分析から得られた基本統計値 (VKS データ)

	N (%)	Max	Min	Mean	SD
Class 1	16 (7.5%)	0.20	-2.06	-0.99	0.54
Class 2	10 (4.7%)	2.04	-1.83	-0.08	1.41
Class 3	48 (22.5%)	1.62	-0.81	0.26	0.51
Class 4	61 (28.6%)	3.12	0.13	1.90	0.62
Class 5	78 (36.6%)	5.83	0.95	2.73	0.89

以下は、Jiao, et al. (2011, pp. 520-522) の中で詳しく説明されているクラス間の分割点を決定する計算手続の要点をまとめたものである。ここでは 5 クラスの分析の Class 3 と Class 4 の分割点計算を例に説明を行う。

$$\ln\left(\frac{W_2\sigma_1}{W_1\sigma_2}\right) = \frac{(x-\mu_2)^2}{2{\sigma_2}^2} - \frac{(x-\mu_1)^2}{2{\sigma_1}^2}$$

W_1: 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$_{(1)}$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$_{(1)}$に隣接するより平均値のより高いクラス$_{(2)}$に所属する受験者数の全体における割合

σ_1:平均値がより低いクラス$_{(1)}$の得点の標準偏差

σ_2:平均値がより高いクラス$_{(2)}$の得点の標準偏差

μ_1:クラス$_{(1)}$の得点の平均

μ_2:クラス$_{(2)}$の得点の平均

Class 3 と Class 4 の分割点の計算

$$\ln\left(\frac{(0.2864 \times 0.5081)}{(0.2254 \times 0.6219)}\right) = \frac{(x-1.8966)^2}{2 \times 0.6219^2} - \frac{(x-0.2634)^2}{2 \times 0.5081^2}$$

この二次方程式を解くと、C3 と C4 の平均値の間に位置することが多い値(x_1)と C3 正規曲線と C4 正規曲線の左側の裾で交わる地点(x_2)の値が算出されるが、x_1 を規準設定の分割点と見なすことができる。

$$x_1 = \frac{-(-1.5512) - \sqrt{-1.5512^2 - 4 \times (-0.2572) \times 1.7889}}{2 \times (-0.2572)} = 0.99053$$

$$x_2 = \frac{-(-1.5512) + \sqrt{-1.5512^2 - 4 \times (-0.2572) \times 1.7889}}{2 \times (-0.2572)} = -7.02192$$

表 1 のクラスの所属数の比率（もしくは素数）を縦軸、横軸に能力値が標示されるように、正規分布曲線の公式と Excel の機能を使って分布の形状を視覚化すると、図 1 のようなグラフになった。C3 の平均値 (0.26) と C4 の平均値 (1.96) の間に交点 x_1(0.9905) が位置していることが示されている。

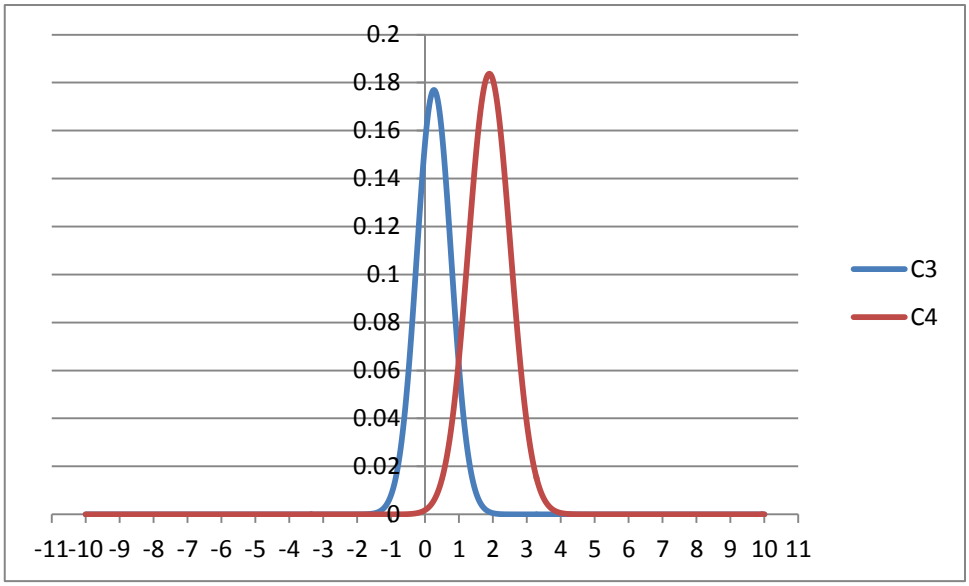


図 1　C 3 と C4 の正規分布曲線 (VKS データ)

表 2 は各クラス間の分割点の計算結果を示したものである。C3 と C4、C4 と C5 の分割点（x_1）は各クラスの平均値の間に位置し、クラスの平均能力値が高くなるにつれて、交差地点（交点 x_1）の値も高くなるが、C1 と C2、C2 と C3 の交点(x_1)の値は、両クラスの平均値の間にはなく、C2 と C3 の交点(x_1)に至っては、平均値が最も低い C1 と C2 の分割点よりもさらに低い値を示していることが確認できる。図 2 からは、C2 の能力分布の幅が大きく、C3 との適切な分割点設定ができない状況が視覚的に示されている。

表 2　5 クラスの MRM 分析から得られた分割点(x_1) の値　（VKS データ）

	Class 1 / Class 2	Class 2 / Class 3	Class 3 / Class 4	Class 4/ Class 5
Cut score (x_1)	-0.0763	-0.9327	0.9905	2.3120

このテストデータからは、C1 と C2、C2 と C3 の分割点を適切に設定することはでき

ず、MRM を使った今回の分析は、法月 (2014) の RM-LR 法の分析とは、異なる結果となった。

クラス数を減らして、4～2 クラスで同様の分析を行ったところ、4 クラス分析では、C1-C2 間の分割点がそれぞれの平均値の間に位置せず、C3-C4 間では両正規曲線が交差する地点が存在しない状況を示す結果となったが、3 クラス、2 クラスの分析では、図 1 のような分割点設定がいずれのクラス間でも可能であることが確認できた。

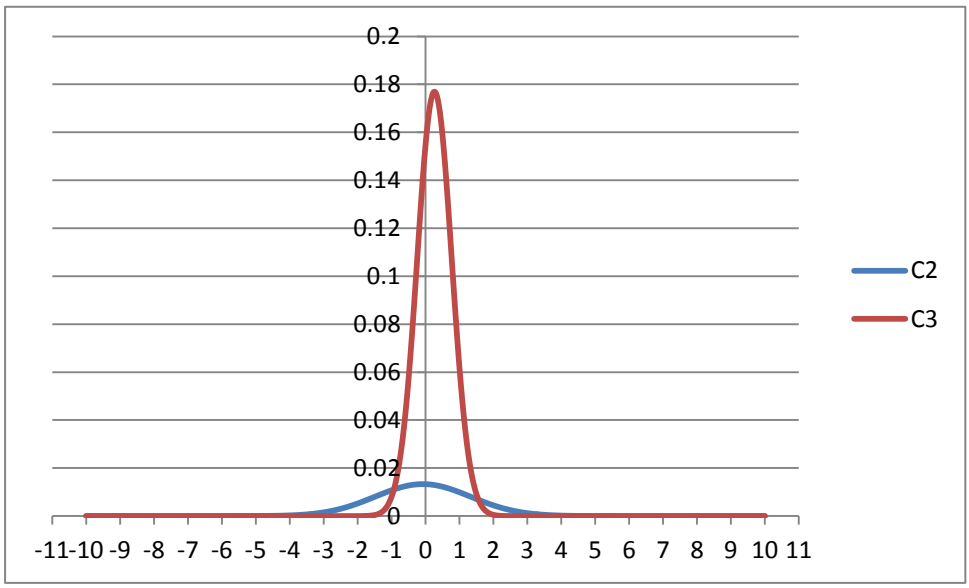


図 2　C2 と C3 の正規分布曲線

1.3-2. 大友 (2013b) との比較

大友 (2013a) は、項目応答理論を使って、Wright and Stone (1979) の Knox Cube Test (以下、KCT データ) 各項目の難易度、識別度、及び正答確率が .67 になる能力推定値を算出し、それぞれの指標がもっとも大きな変化を示す共通項目に分割点を定めているが、大友 (2013b) において同データを RM-LR 法で分析した結果、同じ分割点決定に至ったことが報告されている。そこで、同データを 2 節で述べた MRM 法の 2 クラス分析を使って同様の分割点設定を行うことが可能か検証することとした。

表 3　2 クラスの MRM 分析から得られた基本統計値 と分割点設定結果 (KCT データ)

	N (%)	Max	Min	Mean	SD	Cut score
Class 1	26(76.5%)	5.10	-6.19	0.57	2.72	N/A
Class 2	8(23.5%)	0.62	-3.41	-1.31	1.58	

表 3 は 2 クラス分析のクラスの基礎統計量と両クラスの分割点の計算結果を示している。両クラスの分布の形状は、所属受験者数（%）、平均、最高値、最低値、標準偏差のいず

れの観点からもかなり異なっていることが想像できる。Excel の表計算で x_1、x_2 ともにエラーメッセージ（#NUM!）が標示されたが、これはいずれの計算においても平方根内がマイナスになるため、計算不能であることを意味している。図 3 からも、2 クラスの正規分布曲線がまったく交わっていない状況が明白に示されている。

VKS データの分析においては、法月 (2014) が行ったような 5 ランクの分析は同クラス数の MRM 分析では機能しなかったものの、クラス数を減らした 2、3 クラスの分析では規準設定を行うことが可能であることが示された。一方、KCT データの分析においては、クラス数が少なくても分割点を設定することはできなかった。MRM 法に比べると法月 (2014) や大友 (2013b) の RM-LR 法には、正規分布の交点のような絶対的な規準はなく、複数の分割点候補の中から現実的なニーズを鑑みて柔軟な選択を行うことが許容される。MRM 法についても異なる手順でクラス間を分割することができないか検討する価値があるだろう。

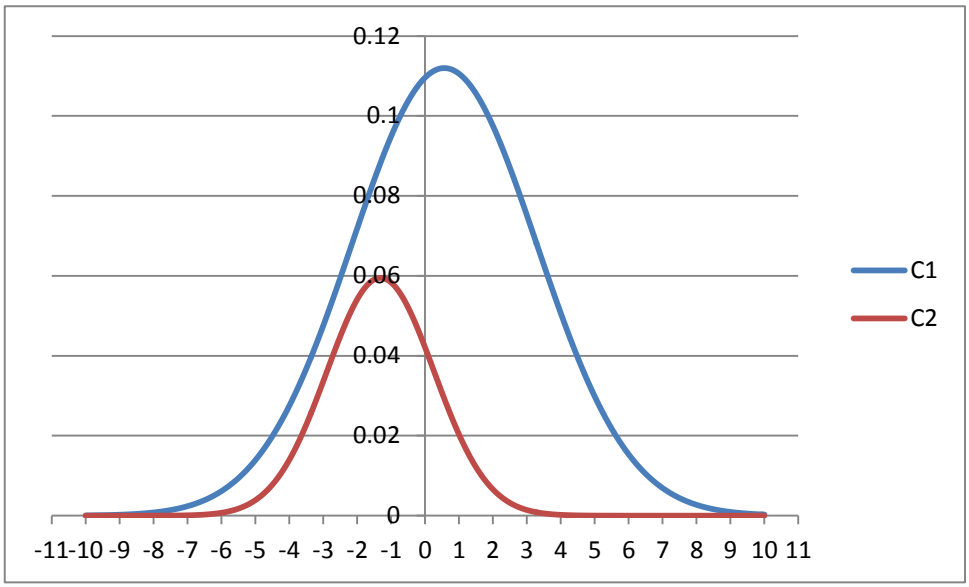


図 3　C1 と C2 の正規分布曲線 (KCT データ)

1. 4. MRM 分析の今後の課題

MRM 法は現段階では一目瞭然の規準設定を導く手法とは言えないが、今後もその様々な側面について研究を進め、客観的で、なおかつ説得力のある規準設定のあり方を模索していく意義がある。

今回の MRM 分析は、VKS データにおいては 5 クラス、4 クラス分析では機能しなかったが、クラス数を減らした 3 クラス、2 クラス分析では適切な分割点を算出することが

できた。一方、KCT データにおいては、2 クラス分析においても分割点の候補になる正規曲線の交点すら得られない結果となった。

考えられる要因としては、VKS データについては、クラス数が多くなると各クラスの所属数が相対的に減り、5 クラス分析の C2 のように能力分布が広くなり、特定の能力層で構成されていると考えられる他のクラスと比較が困難になっている可能性がある。

KCT データについては、大友 (2013a, 2013b) の手順と同様に、全問不正解の受験者 1 名と全員正解 3 問及び全問不正解 1 問の項目を除去した 34 名、14 項目の極めて小さなデータの分析であったが、通常大きなサンプル数を必要とする MRM 分析には十分に適応しなかったのかもしれない。

また、今回の分析では、MRM 法を他の統計的解決法と規準設定の観点から比較することが最も重要な研究テーマであったが、法月 (2014) や大友 (2013b) の RM-LR 法や、大友(2013a) の項目応答理論に基づく解決法は、いずれも正規曲線の交点から絶対的な分割点の数値を導くような方法ではなかったため、結果を単純に比較することはできない。

WINMIRA 2001 は、ラッシュモデルの能力推定値とクラス分類を同時に算出してくれる。クラスについては能力平均値の昇順に配列を変える必要があるが、その比較的単純な作業を済ませた後は RM-LR 法と同じような手続きで分割点を設定することも可能ではないかと考えられる。このような現状の問題点を認識した上で、以下の３つの研究課題を検証する価値がある。

課題 1　MRM 法で小グループデータの分析を行うことは適切ではないか。適切ではない場合、他のソフトウェアの使用、シミュレーション分析を含めて、どのような対応処置が可能か。

課題 2　MRM 分析後に行う規準設定手続きを、隣接するクラスの正規分布曲線の交点決定とは異なる条件、あるいは類似の柔軟な条件で行うことは可能か。

課題 3　MRM 法と RM-LR 法とのより合理的な比較検証は可能か。

法月 (2014) が分析した VKS データには、テスト問題以外に 4 段階の自己評価 （語彙認識度） 問題があるが、このような評定スコアの分析については、英検の二次面接 （スピーキング） テストや現在 1 級、準 1 級で実施されていて、2016 年度第 1 回試験から 2 級でも導入されることが決定したライティングテスト （日本英語検定協会, 2015） の分析にも応用できるであろう。MRM は、規準設定の観点から評定スコアの分析にどのように活用することができるだろうか。また、MRM に基づく等化手続きを経て、VKS 1 や VKS2、英検準 1 級と 2 級の問題、あるいは英検 2015 年度第 1 回問題と第 2 回問題を同一尺度上に位置づけ、安定した規準設定を行うことは可能だろうか。このような見地から、上記の課題 1～3 が十分に検証された後には、以下の 2 つの課題も追究していくことが大いに期待できる。

課題４　MRM 法を使って言語テストの多値 (polytomous) データの規準設定を効果的に行うことは可能か。

課題５　MRM 分析を使って現実的な等化手続きに基づく規準設定を行うことは可能か。

参考文献

American Educational Research Association, American Psychological Association and National Council On Measurement in Education. (2014). *Standards for educational psychological testing.* American Educational Research Association.

Baghaei, P. and Carstensen, C. H. (2013). 'Fitting the mixed Rasch model to a reading comprehension test: Identifying reader types' in *Practical Assessment, Research & Evaluation., Vol. 18, No. 5.*

Bond, T.G., Fox, C.M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mhwah, Nj: Erlbaum.

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)

Davier, M. (1997). 'WINMIRA – program description and recent enhancement' *in Method of Psychological Research Online, 1997, Vol. 2, No. 2.* Pabst Science Publishers.

Griffin, P., McGaw B. & E. Care. (eds.). (2012). *Assessment and teaching of 21st century skills.* Springer.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory.* Sage Publications, Inc.

Jiao, H., Lissitz, R.W., Macready, G., Wang, S., and Liang, S. (2011). 'Exploring levels of performance using the mixture Rasch model for standard setting' in *Psychological Test and Assessment Modeling,* Vol. 53, 2011 (4). (pp. 499-522).

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.

Lazarsfeld, P.F. (1950). Chapter 10, 11 in Stouffer et al. *Measurement and Prediction.* Princeton Univ. Press.

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

Lissitz, R.W. (2013). 'Standard setting: past, present, and perhaps future'. In M. Simon, K.

Ercikan & M. Rousseau (Eds.) *Improving large-scale assessment in education: Theory, issues, and practice*. (pp.154-174). New York: Routledge.

Lord, F. M. (1952). *A theory of test scores* (*Psychometric Monograph No.7*). Psycho- metric Society.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Addison-Wesley.

Muthén, L.K., & Muthén, B.O. (2012). *Mplus User's Guide. Seventh Edition.* Los Angeles, CA: Muthén & Muthén.

Preinerstorfer, D. and Formann, A. K. (2012). 'Parameter recovery and model selection in mixed Rasch models'. *British Journal of Mathematical and Statistical Psychology, 65,* (pp. 251-262).

Rost, J. (1996). 'Logistic Mixture Models' in Linden, J. & R. K. Hambleton. (1996). *Handbook of modern item response theory*. Springer. (pp. 449 – 463).

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.

Templin J. & Jiao, H. (2012). 'Applying model-based approaches to identify performance categories'. in Cizek, G. J. (ed). (2012). *Setting performance standards, second edition.* (pp. 379 – 397).

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

von Davier, M. (2014). 'Proficiency Levels and Standards from a Latent Variable Modeling Perspective' (Workshop at the Research Center for Psychological and Educational Testing, Taipei, Taiwan, January 16-17, 2014).

Wright, B.D., & Stone, M.H. (1979). *Best test design.* Chicago: MESA Press.

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. （1996）. 『項目応答理論入門』. 大修館書店.

大友賢二 研究代表・渡部良典 研究副代表. (2012). 『言語テストの規準設定 報告書第1号』. 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・渡部良典 研究副代表. (2013). 『言語テストの規準設定 報告書第2号』. 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. (2013a). 「予備調査：CITO variation on the bookmark method」. 『言語テストの規準設定 報告書第２号』. 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. 1-38).

大友賢二. (2013b). 「英語教育とテスト：第二言語習得における規準設定をめぐって」.

『第７回日本テスト学会賞記念講演』. 東京：成蹊大学.
大友賢二 研究代表・渡部良典 研究副代表. (2014). 『言語テストの規準設定　報告書第３号』. 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. (2015). 『ICT 等を活用した評価についての調査・研究　報告書』. 公益財団法人英語検定協会英語教育センター委託研究.
岡本昭・中嶽治磨•村瀬隆二・山本輝夫 編. （昭和４４年）. 『教育における統計事典』. 三晃書房.
日本英語検定協会. (2015 年). 「実用英語技能検定『2 級』ライティング導入のお知らせ」. Retrieved from https://www.eiken.or.jp/eiken/info/2015/pdf/20150715_pressrelease_writing2.pdf.
法月　健. （201３）. 「受容語彙力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」. 『言語テストの規準設定：報告書第２号』. 英検： 英語教育研究センター委託研究.
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」. 『言語テストの規準設定：報告書第３号』. 英検： 英語教育研究センター委託研究.
東 洋・梅本襄夫・芝祐順・梶田叡一編. （1988）. 『現代教育評価事典』. 金子書房.
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/.
依田新 監修. （1977）. 『新・教育心理学事典』. 金子書房.

Kagi on line store: http://order.kagi.com/cgi-bin/store.cgi?storeID=1HN&&
Rasch Measurement Analysis Software Directory: http://www.rasch.org/software.htm
Rupp, A. A. (2009). "Software for Calibrating Diagnostic Classification Models: An Overview of the Current State-of-the-Art" *retrieved at* http://www.education.umd.edu/EDMS/fac/Rupp/AERA-SIG%20Software%20Symposium%20(Handout%20Package).pdf, on Feb. 6, 2015.
WINMIRA manual: http://208.76.84.140/~svfklumu/wmira/winmiramanual.pdf
WINMIRA 2001: http://208.76.84.140/~svfklumu/wmira/index.html

荘島宏二郎. (n.d.). 「エグザメトリカ」 http://www.rd.dnc.ac.jp/~shojima/exmk/jindex.htm

Mixture Rasch Model による 英語能力の規準設定

2015年度　進捗状況報告
2015年10月16日

大友賢二・中村洋一・法月　健

2.1. 規準設定に関する 方法と統計処理

1. Methods that involve review of test items and scoring rubrics
2. Methods that involve review of candidates
3. Methods that involve looking at candidate work
4. Methods that involve panelist review of score profiles

(Hambleton and Pitoniak (2006))

Rasch Model (RM)
Latent Class Analysis (LCA)
Mixtue Rasch Model (MRM)

2.1-1. ラッシュ・モデル (Rasch Model：RM：IRT)

(1) 受験者の能力は、解答した項目群とは独立に定義される。
(2) 項目の困難度は、受験者集団とは独立に定義される。
(3) 項目の困難度と受験者の能力値とが同一の尺度上で表現されるので、当該受験者に提示した項目が適切であったかどうかの判断が容易にできる。

2.1-2. 潜在クラス分析 （latent class analysis: LCA）

潜在クラス分析では、被検査者群が、潜在的で直接観察できない層（潜在クラスという）に層別されていて、各層ごとに、調査や監査の項目への応答率が異なるものと仮定する。被験者全体としての各項目への反応のデータを用いて、潜在クラスの個数、潜在クラスの相対的な大きさ、潜在クラス別の各項目への応答率を推定しようとする手法である。（生澤, 1988, p.370)

2.1-3. 混合ラッシュモデル (Mixture Rasch Model: MRM)

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 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 (Lissitz, 2013, p. 170) 統計的解決モデルと言える。

2.2. MRMの利点

- 現在、ほかで用いられている規準設定法においては、事前に明記される能力水準に関する統計的な妥当性検討はなされていない。
- 最低能力とか境界線上の受験者を確認できる。
- テスト項目やテスト受験者の部分集合を用いて分割点の相互の妥当性検討は容易にできる。
- この手順は、テスト中心法にも、受験者中心法にも使える。
- （Jiao, et al. 2011, pp.514-515)

2.3. MRMを使った規準設定手順

$$\ln\left(\frac{W_2\sigma_1}{W_1\sigma_2}\right) = \frac{(x-\mu_2)^2}{2{\sigma_2}^2} - \frac{(x-\mu_1)^2}{2{\sigma_1}^2}$$

W_1: 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$_{(1)}$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$_{(1)}$に隣接するより平均値のより高いクラス$_{(2)}$に所属する受験者数の全体における割合
σ_1:平均値がより低いクラス$_{(1)}$の得点の標準偏差
σ_2:平均値がより高いクラス$_{(2)}$の得点の標準偏差
μ_1:クラス$_{(1)}$の得点の平均
μ_2:クラス$_{(2)}$の得点の平均

Jiao, et al. (2011, p.521)

2.3-1. WINMIRA2001

- The models described in the previous sections can be computed by means of the Windows-program WIN-MIRA (MIxed RAsch model: von Davier (1995)). (Rost (1996, p.457))
- Matthias von Davier is a Senior Research Scientist in the Research & Development Division at Educational Testing Service.(von Davier & Carstensen (2007)
- WINMIRA is a software for analyses with a variety of discrete mixture distribution models for dichotomous and polytomous categorical data. This soft ware can be used for the Rasch Model, the Latent Class Analysis, the Mixed Rash Model, and the Hybrid Model. (von Davier, 2000, 2001)

2.3-2. WINMIRA による分析

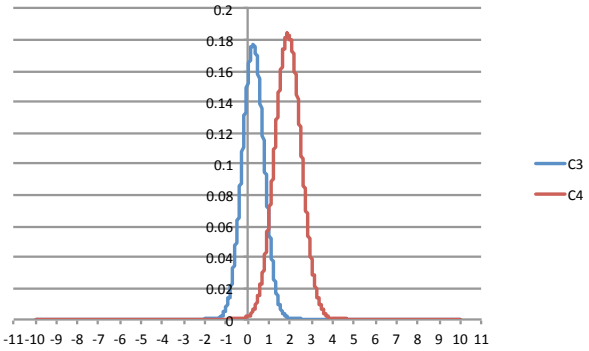


図1 C3とC4の正規分布曲線 (VKSデータ)

表2　5クラスの MRM 分析から得られた分割点(x_i) の値（VKS データ）

	Class 1 / Class 2	Class 2 / Class 3	Class 3 / Class 4	Class 4/ Class 5
Cut score (x_i)	-0.0763	-0.9327	0.9905	2.3120

2.4. MRM分析の今後の課題

- 課題1　小グループデータの分析
- 課題2　MRM分析後に行う規準設定手続きの条件
- 課題3　MRM法とRM-LR法とのより合理的な比較検証
- 課題4　多値 (polytomous) データの規準設定
- 課題5　現実的な等化手続きに基づく規準設定

参考文献

- Hambleton, R.K. & Pitoniak, M.J. (2006). Setting Performance Standards.
- In R.L. Brennan (Ed.) *Educational Measurement: Fourth Edition*
- (433-470). American Council on Education and Praeger Publishers.
- Jiao, H., Lissitz, R.W., Macready,G., Wang, S., & Liang, S. (2011). Exploring
- levels of performance using the mixture Rasch model for standard
- setting. In *Psychological Test and Assessment Modeling*, Vol.53.,
- 499-522.
- Lissitz, R. W. (2013). Standard Setting: Past, Present and Perhaps Future. In
- M.Simon, K. Ericikan, & M. Rousseau (Eds.). *Improving Large-Scale*
- *Assessment in Education*. (154-174). Routledge Taylor & Francis Group.
- Rost, J. (1966). Logistic Mixture Models. In W. J. van der Linden & R. K.
- Hambleton (Eds.*) Handbook of Modern Item Response Theory*.
- 449-463. Springer.

参考文献

- Templin, J. & Jiao, H. (2012). Applying model-based approaches to identify
- performance categories. In Cizek, G.T. (ed.) Setting performance standards:
- second edition (pp.379-397).Routledge.
- Von Davier, M. (2000, 2019) WINMIRA 2001
  Von Davier, M. & Carstensen, C.H. (Eds.)(2007). *Multivariate and Mixture*
- *Distribution Rasch Models*. Springer.

- 生澤雅夫(1988). 潜在構造分析. 東、梅本、芝、梶田（編）.『現代教育評
- 価事典』金子書房、p.370.
- 法月 健 (2012). VKS (Vocabuary Knowledge Survey: VKS).

- ＊＊＊＊＊

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

３．１．受容語彙力テストの分析と今後の課題

法月　健

１．2015 年度の研究結果のまとめ：中間報告の要旨と補足

2015 年度の研究中間報告（大友・中村・法月、2015）において、法月は、「言語テストの規準設定報告書　3」中の法月 (2014) が使用した英検の級別に対応した 2 種類の受容語彙力テスト 50 問・自己評価 50 問の評価システム(VKS1 & VKS2) のうち、4 級～準 1 級に対応した VKS1 のテスト部門 50 問の 213 名のデータ（以降、VKS データ）について、２～５クラスの MRM 分析を行い、法月 (2014) で採択したラッシュモデル(RM)と潜在ランク(LR) 理論を併用した規準設定法（以降、RM-LR 法）の結果と比較を行った。

MRM 分析には、WINMIRA 2001（von Davier, 2001）を使用し、Jiao, Lissitz, Macready, Wang & Liang (2011, 499-522) が議論した確率密度関数の概念に基づき、下記の数式から 2 次方程式の解の公式を導き、C1～C5 の隣接するクラス（階層）間で正規分布曲線が交わる地点を計算する分割点設定方法を適用した。

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

W_1: 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$_{(1)}$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$_{(1)}$に隣接するより平均値のより高いクラス$_{(2)}$に所属する受験者数の全体における割合

σ_1:平均値がより低いクラス$_{(1)}$の得点の標準偏差

σ_2:平均値がより高いクラス$_{(2)}$の得点の標準偏差

μ_1:クラス$_{(1)}$の得点の平均

μ_2:クラス$_{(2)}$の得点の平均

分析の結果、C3 と C4、C4 と C5 の分割点（x_1）は各クラスの平均値の間に位置し、クラスの平均能力値が高くなるにつれて、交差地点（交点 x_1）の値も高くなるが、C1 と C2、C2 と C3 の交点(x_1)の値は、両クラスの平均値の間にはなく、C2 と C3 の交点(x_1)に至っては、平均値が最も低い C1 と C2 の分割点よりもさらに低い値となった。また、C2 の能力

分布の幅が大きく、より平均値の高い C3 の最高値と最低値が含有される状況となった。今回の５クラス MRM 分析においては、C1 と C2、C2 と C3 の分割点を適切に設定することはできず、同じ VKS データを使用した法月 (2014) の RM-LR 法の分析とは、異なる結果となった。

複数の潜在クラスを仮定しない標準的な RM と異なり、MRM では同じ素点であっても異なる潜在クラスに位置付けられることもあり、能力パラメタ―値も一定ではない。たとえばデータ中に８名の素点 41 点（正答率 82 点）を記録した受験者のうち、7 名は最も平均値が高い C5（78 名で構成）に所属し、MRM 能力パラメタ―値は 3.150 （受験者平均 1.524）であったが、残り 1 名は平均値が 2 番目に低い C2（10 名で構成）に所属し、能力値は2.043だった。C2には素点41点の受験者1名以外に40点(MRM1.777)の受験者1名 (C5 に 6 名) など高得点の受験者が所属する一方で、素点 13 点(MRM -1.827)、15 点 (MRM -1.577) のように得点の低い受験者も含まれている。わずか 10 名で構成される C2 は、このように特定の能力層を示すクラスを形成しているとは言えず、その能力パラメタ―値の分布を概観するだけで、図 1 の様に隣接する C1 や C3 のクラスとの間に意味のある分割点が設定できない状況にあることを、容易に理解することができる。

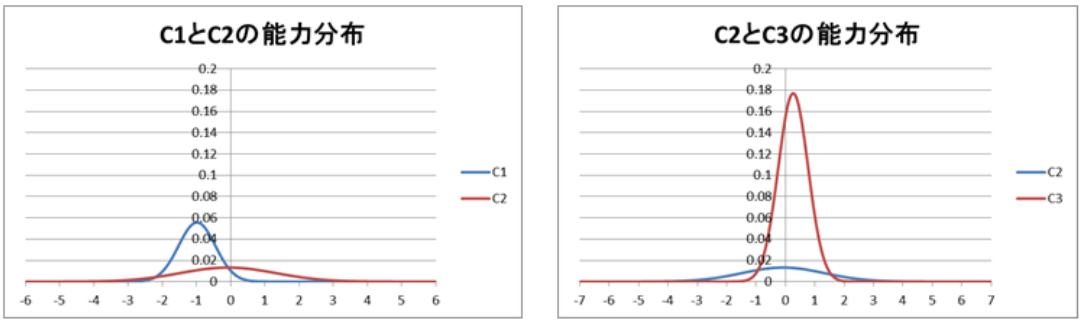


図１　５クラス分析における C2 と隣接する潜在クラスの正規分布曲線

５クラス分析の C2 のように極端にサイズの小さいクラスが形成される可能性を軽減するため、クラス数を減らして、４～２クラスで同様の分析を行ったところ、４クラス分析では、C1-C2 間の分割点がそれぞれの平均値の間に位置せず、C3-C4 間では両正規曲線の交点が存在しない状況を示す結果となったが、３クラス（図２参照）、２クラスの分析では、分割点設定がいずれのクラス間でも可能であることが確認できた。

次に、全問不正解の受験者 1 名と全員正解 3 問及び全問不正解 1 問の項目を除去した 34 名、14 項目の Wright and Stone (1979) の Knox Cube Test (以下、KCT) データを使用して、MRM 法の２クラス分析を行った。　大友 (2013a) では、各項目の難易度、識別度、及び正答確率が.67 になる能力推定値を算出し、それぞれの指標がもっとも大きな変化を示す共通項目に分割点を定めているが、大友　(2013b) において同データを RM-LR 法で分析した

結果、同じ分割点決定に至ったことが報告されている。

KCT データの MRM 分析の結果、2 次方程式の解の公式の計算の過程で、平方根の中がマイナスになるため、計算不能となり、グラフ化すると、２クラスの密度関数が交わることなく分布している状況が示された（大友・中村・法月、2015）。

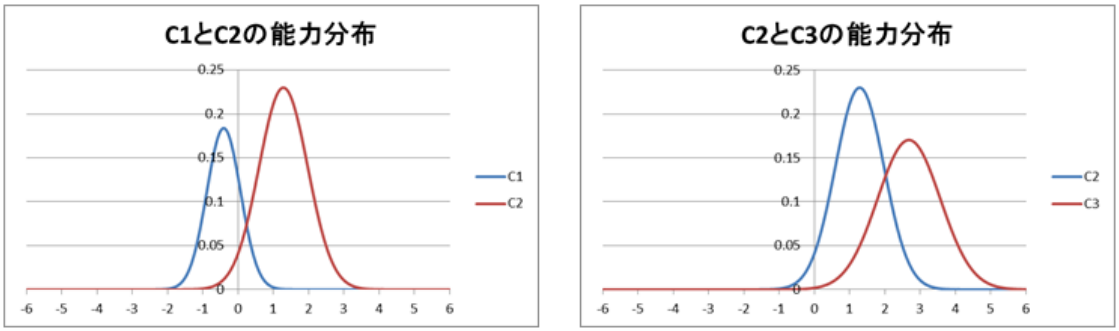


図２　３クラス分析の分割点設定

VKS データの５クラス MRM 分析は、法月 (2014) が行った同クラス数の RM-LR 分析とは異なり、十分に機能しなかったが、クラス数を減らした２、３クラスの分析では規準設定を行うことが可能であることが示された。その一方で、KCT データの分析においては、クラス数が少なくても分割点を設定することはできなかった。MRM 法に比べると法月 (2014) や大友(2013b) の RM-LR 法には、正規分布の交点のような絶対的な規準がなく、複数の分割点候補の中から現実的なニーズを鑑みて柔軟な選択を行うことが許容される。MRM 法についても異なる手順でクラス間を分割することができないか検討する価値があるだろう。

２．今後の課題

大友他 (2015）では、以下の５つの研究課題が提案された。

課題１　MRM 法で小グループデータの分析を行うことは適切ではないか。適切ではない場合、他のソフトウェアの使用、シミュレーション分析を含めて、どのような対応処置が可能か。

課題２　MRM 分析後に行う規準設定手続きを、隣接するクラスの正規分布曲線の交点決定とは異なる条件、あるいは類似の柔軟な条件で行うことは可能か。

課題３　MRM 法と RM-LR 法とのより合理的な比較検証は可能か。

課題４　MRM 法を使って言語テストの多値 (polytomous) データの規準設定を効果的に行うことは可能か。

課題５　MRM 分析を使って現実的な等化手続きに基づく規準設定を行うことは可能か。

いずれも今後取り組むべき重要な課題であるが、まず、課題１～３を関連付けながら、以下で述べる視点を軸にして、検証を進めたい。

３．小グループデータの分析

Jiao et al. (2011) は、確率密度関数を応用した MRM 法の分析を正確に実施するには 1 万人規模のデータが必要であることを示唆しているが、34 名の KCT や VKS のような数百名程度のデータ分析では大きな測定誤差を生じ、特に成員が少ないクラスを含む分析には十分に適応しなかった可能性がある。対応策として、実データを基にシミュレーションデータを生成することで、大規模データが有する必要条件を満たす分析を行うことが期待できるかもしれない。これには、Mplus (Muthen & Muthen, 2012) や mRm ((Preinerstofer & Forman, 2012) の様なソフトウェアの使用が考えられるが、相応のプログラミング技術を有することが効果的に実践するための最低条件と言える。また実データ（基データ）の特徴が詳細に記述されたシミュレーションデータを活用した規準設定の研究事例が見当たらず、仮に理論的にそのようなシミュレーション分析が可能であっても、現実的な限られた人的・物理的資源の中で、どの程度に運用可能なのかは不明である。シミュレーション分析への発展も視野に入れつつ、シミュレーションを伴わない比較的少人数のデータにも対応する安定した分析法が他にないか検討する必要もあるだろう。

４．MRM 分析後に行う規準設定手続きの条件

Kreiner (2007) は認知症の判別検査 (1,147 名) の分割点を MRM と関連の統計手法を使って、分析している。まず、得点分布を指定しない MRM 分析によって得られる赤池情報量規準 (Akaike’s information criterion: AIC) の値の比較から、クラス (階層) 数を２つに分割すべきと判定し、次に、①Andersen (1973) の条件付き尤度比 (conditional likelihood ratio) 、②Benjamini & Hochberg (1995) 等の局所的均質性 (local homogeneity)、③MRM のクラス間の発生頻度期待値 (expected frequency) 比の分析を基に、分割点の位置（素点）の候補を 2 点に絞り込んでいる。①と②の分析と MRM に基づく③の分析が、異なる 2 点の分割点候補を支持する結果となったが、診断結果が陰性で実際に発症していない「真の陰性 (true negative)」の確率 (specificity) が高い地点については、両候補の数値が高く、大差がないため、診断結果が陽性で実際に発症している「真の陽性 (true positive)」の確率 (sensitivity) が顕著に高い値を示す①と②の結果が支持する地点を、最終的に妥当な分割点と結論付けている。

Kreiner (2007) の分析結果は、242 人のサブサンプルに対しても有効であったことが示されているため、その分析手法が VKS データのような数百名程度の分析にも適用できる可

能性があると言える。③の手法は確率密度関数の交点を基準とする Jiao et al. (2011) や大友他 (2015) の概念に通じる同様の解析法と考えることができるが、分析結果から、MRM 法の分析結果が常に絶対的とは言えず、異なる分析法を併用することで、より複合的な見地から分割点設定を検証することが可能であると解釈して良いだろう。

同様の研究として、Kreiner, Hansen & Hansen (2006) も類似の複合的な分割点設定の手法を使用しているが、MRM については WINMIRA 2001 の結果データから算出や抽出が可能な各得点における隣接するクラス間の条件付き項目応答確率 (the conditional probabilities of item responses) と混合比率 (mixing proportions) の割合の対比から分割点の位置を探る方法を提案しており、こちらも MRM を応用した分析法候補の一つとして注目される。

Kreiner (2007) と Kreiner et al. (2006) の MRM 分析手続きに従って、VKS データを 2 分割する地点を探ったところ、いずれも素点の 29 点と 30 点の間に分割点を設定すべきであることが示唆される結果となった。大友他 (2015)では Jiao et al. (2011) の方法に従って MRM 能力パラメタ―値を使って分析したが、素点の平均と標準偏差の値を基に計算を行うと、他の分析結果に符号する 29.964 の値を示した。

素点で標示された分割点は、一般のテスト利害関係者にとって解釈しやすく、有用で利用しやすい情報だと思われるが、RM と異なり MRM の能力パラメタ―値は素点と一対の関係にはない。実際、MRM 能力パラメタ―値による分割点は 0.933 であったが、この値は C2 の素点では 29 点と 30 点の間に位置する一方で、C1 の素点では 32 点と 33 点の間に位置することになり、にわかに素点に換算することはできない。

また２分割の状況では Kreiner (2007)、Kreiner et al. (2006)、大友他 (2015) の異なる方法で一致した判定結果を残したが、より多くの潜在クラス数を仮定する場合も同様の結果が得られるだろうか。５分割のデータを素点で分析してみると、C4-C5 以外のクラス間には適切な分割点を見出すことができず、素点による分析に限界があることが示唆される。

分割点設定における素点と MRM 能力パラメタ―値の扱いについては、今後の研究課題の中で検証を続けながら、認識を深めていくことが求められるが、今後、Kreiner (2007) や Kreiner et al. (2006) の分析で活用された①Andersen (1973) の条件付き尤度比、②Benjamini & Hochberg (1995) の局所的均質性等との併用でさらに分析の客観性を高め、多角的な統計的判定法の確立に向けて、検証を続けていきたい。

５．MRM 法と RM-LR 法とのより合理的な比較検証

MRM 法と RM-LR 法を比較した研究は筆者の知る限り皆無と言え、両者の分析の背景を成す理念に相反する側面があることは否めないが、提示される分析データの数理的特徴には類似性も多く、比較検証を行う価値は高いものと考えられる。

法月 (2014) は、RM-LR 法を使って、望まれる規準に達していない学習者が自分の習熟度よりも高いグループに誤って配置される可能性を極力減らす「真の陽性」と望まれる規

準に達している学習者が自分の習熟度よりも低いグループに誤って配置される可能性を極力減らす「真の陰性」の両極の観点から、現実的で柔軟な分割点設定法のあり方について議論したが、MRM 法に基づく方法にも、このような概念を適用することは可能であろうか。

Kreiner (2007) の扱った医学データと異なり、「真の陽性」、「真の陰性」を客観的に判別することは、残念ながら英語能力については決して容易とは言えない。VKS では各問題の理解度を解答直後の問において 4 段階で自己診断するシステムを設けているが、このような指標を MRM や LRT の指標と比較することで、素点からは判断できない受験者能力の特徴を探ることができるかもしれない。

たとえば、問 31 に正解した受験者が問 32 で「１」（見たこともないし、意味も分からない単語）、もしくは「２」（見たことはあるが意味は分からない単語）を選んでいる場合は、問 31 における理解度と関係なく、当て推量や誤った類推から偶然に選んだ選択肢が正解だった可能性がきわめて高いことを意味している。

このように理解度が低いながらも全般に正解率が高い受験者については、MRM や LRT の指標にどのような影響が出るのだろうか。たとえば、素点で 41 点を記録しながら唯一 MRM の５クラス分析で能力平均値が下から 2 番目に低い C2 に位置付けられた受験者には、LRT の５ランク分析データ (法月 2014) において、41 点以上の受験者 28 名中唯一トップのランク５でない「４」のランクが付与された。この受験者の理解度平均は上位 28 名中 25 位タイであり、理解度の相対的な低さが MRM の潜在クラスや LRT の潜在ランクの決定に影響が出ていた可能性もある。理解度の分析に対して、MRM 法を使った多値データの分析も視野に入れるならば、将来的に上記の課題４の研究へと発展させていく契機となるだろう。

また RM-LR 法の分析結果を基に、MRM 法と同様の手続きに従って、確率密度関数の交点を計算し、分割点を設定することは可能であろうか。MRM の 5 クラス分析では、C5 と C2 に分かれた素点 41、40 点の受験者に代表されるように、複数の潜在クラスに同じ素点の受験者が広く分布する傾向が強く見られた。一方で、法月 (2014) の RM-LR 法のランク付与結果を見ると、同じ素点でも異なるランクが付与されたり、素点の低い受験者が素点の高い受験者よりも上位のランクに位置付けられることはあったものの、ランク５が１名、４が８名、３が１名となった素点 37 点の受験者を除き、各素点は１つもしくは２つのランクに位置付けられていた。このことから、RM-LR 法では、MRM 法以上に安定したランク間の交点計算ができる可能性も期待できる。もし RM-LR 法と MRM 法で著しく異なる分割点設定が示唆される場合は、結果をどのように解釈するべきかの問題に対峙しなければならないが、それは簡単には解決できない難しい研究課題となるだろう。

このような探究を進めることは、どの単一の統計的解決法が分割点設定法として優れているかを見極めるのではなく、多角的な視点から、より柔軟で合理的な分割点設定の枠組

みを論ずる絶好の機会となる。また、４節で述べた Kreiner (2007) や Kreiner et al. (2006) が議論した複合的な MRM 法、大友（2013b）が提唱した他の項目応答モデル分析との比較検証とも関連付けて有機的な研究の形成につながり、大いに意義があることと言えるだろう。

参考文献

Andersen, E.B. A goodness of fit test for the Rasch model. *Psychometrika, 38,* 123-140.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B, 57,* 289-300.

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.

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.

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

Muthén, L.K., & Muthén, B.O. (2012). *Mplus User's Guide. Seventh Edition.* Los Angeles, CA: Muthén & Muthén.

Preinerstorfer, D. and Formann, A. K. (2012). Parameter recovery and model selection in mixed Rasch models. *British Journal of Mathematical and Statistical Psychology, 65,* 251-262.

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

Wright, B.D., & Stone, M.H. (1979). *Best test design.* Chicago: MESA Press.

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」、『言語テストの規準設定　報告書第３号』、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.77−101).

大友賢二 (2013a). 「予備調査：CITO variation on the bookmark method」　『言語テストの規準設定　報告書第２号』、 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.1-38).

大友賢二 (2013b、１２月). 「英語教育とテスト：第二言語習得における規準設定をめぐって」、　『第７回日本テスト学会賞記念講演』、　東京：成蹊大学.

大友賢二・中村洋一・法月 健 (2015).「英語教育研究センター2015 年度委託研究中間報告：Mixture Rasch Model による英語能力の規準設定」公益財団法人 日本英語検定協会

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

３．２．検討必要性の再確認

中村 洋一

本研究は、公益財団法人日本英語検定協会英語教育センター委託研究として、2012 年より、言語テスト・英語能力の規準設定について研究を続け、従来の規準設定法の整理と併せて、新たな分割点設定の統計的処理の可能性を模索してきた研究 (大友 他, 2012, 2013, 2014, 2015)の延長線上にある。我が国が、新たな英語教育の舵取りを検討し直している今、英語能力の規準設定という未解決の課題の解決に取り組むことは不可欠であるが、必ずしも充分な研究がなされているとは言えない状態である。本稿では、英語能力の規準設定を検討する必要性について再確認し、さらなる研究継続の方向性を提案するものとする。

2006 年に出版された *Handbook of test development* は、10 年を経て、その第 2 版が今年 2016 年に出版された。2nd Edition の Preface は、ここ数年で項目応答理論・一般化可能性理論・妥当性検証の理論の、テスト開発に関する問題への適用といった心理測定研究の進歩が見られたこと、テストの作成、実施、採点、コンピュータ適応型テストに供するコンピュータ技術の進歩が熟成していることをあげている。しかし、一方で、However, many of the activities associated with high-quality test development remained undocumented (p. xv).、The art and science of test development are often learned in testing agencies through on-the-job training, with best practices handed down from master to apprentice. Psychometric theory, to be sure, provides the foundation for all test development activities, but the nexus between theory and practice is not always apparent (p. xvi). と指摘し、2nd Edition 出版の動機を記している。

2006 年版の *The Handbook* では、Ⅱ. Content のパートにおいて、Cizek, G. J. が 10. Standard Setting の項目で、スタンダード・セッティングについて執筆している。2016 年版の 2nd Edition では、執筆者が Cizek & Earnest の二人になり、項目のタイトルも 11. Setting Performance Standards on Test (pp. 212 - 238) と改訂されている。その中で、*Standards for educational psychological testing* の最新版における Standards related to Setting Cut Score の改訂

を引用し (p. 214)、Standard 7.4 の Test documentation should summarize test development procedures, including … the methods for establishing performance cut scores. を新たに加えたことを紹介している。また 2006 年版にはなかった、NCLB Act における K-12 のアセスメントで検討されている Vertically Moderated Standard Setting: VMSS (pp. 232 – 234) にも触れている。2nd Edition のこの項の最後には、Frontiers and Conclusion のひとつとして … a comprehensive set of guidelines does not exist regarding what should be included in a technical report on standard setting – or even a guideline that such a document should be produced. Clearly, it is a professional obligation to document the procedures, findings and limitations of any standard-setting activity, either as a separate report or as a section within broader technical documentation for a testing program. を挙げている。

本研究の今後の課題も、この obligation と同一のものであると考える。つまり、緒についたばかりの、我が国における規準設定の方法論研究において、小さな一歩だとしても、Mixture Rasch Model: MRM の適用可能性に関する研究を手がかりとして、客観的で、なおかつ説得力のある規準設定のあり方をさらに模索していく必要がある。その手始めとして、本研究の中間報告 (大友・中村・法月, 2015) で取り上げられた課題について、今後の方向性を以下に記すこととする。

課題 1　MRM 法で小グループデータの分析を行うことは適切ではないのか。適切でない場合、他のソフトウェアの使用、シミュレーション分析を含めてどのような対応処置が可能か。

本研究のこれまでのデータ分析のプログラムとしては、WINMIRA (von Davier, M. 2001) を使用してきたが、中間報告で示したように、分析にやや問題が生じた。その原因のひとつとしてデータ数が少ないことが考えられるが、その問題の原因追及と、より適切なアルゴリズムや、プログラムの選定を見据えた研究をしていくことが必要である。WINMIRA の他には、Templin & Jiao (2012) が使用した mdltm といったプログラムもある。Hagenaars & McCutcheon (eds.) (2002) には、Appendix C: Selected software として、WINMIRA, PANMARK, LATENTGOLD, LEM, MPLUS, WINLTA, GLIMMIX, DNEWTON といったプログラムのリストが示されている。Rasch Measurement Analysis Software Directory (http://www.rasch.org/software.htm) といったウェッブページもある。また、Frick, Strobl, Leisch & Zeileis の psychomix のような R 上で稼働するプログラムなども開発されている。このようなプログラムを検証することにより、実際の、あるいはシミュレーションにより作成されたデータを用いて、より適切なアルゴリズムとプログラムの適用方法について検討することが必要である。

課題 2　MRM 分析後に行う規準設定手続きを、隣接するクラスの正規分布曲線の交点決定とは異なる条件、あるいは類似の柔軟な条件で行うことは可能か。

課題 3　MRM 法と RM-LR 法とのより合理的な比較検証は可能か。

大友・中村・法月 (2015) は、Jiao et al. (2011) に則った方法を中心に検討した。そして、課題として「法月 (2014) や大友 (2013b) の RM-LR 法や、大友(2013a) の項目応答理論に基づく解決法は、いずれも正規曲線の交点から絶対的な分割点の数値を導くような方法ではなかったため、結果を単純に比較することはできない」ことをあげた。Cummings & Petscher (eds.) (2016) では、Latent Class Analysis: LCA の枠組みを使用して Cut Score を設定する方法論についていくつかの例を提示している。また、Huynh & Scheneider (2005) や、Ferrara, , Hohnson, & Chen (2005) のように、Vertically Moderated Standard Setting: VMSS の方法論を検討する文献もある。様々な文献研究とあわせて、課題 1 の検討も進め、規準設定の方法論に関する研究を蓄積していく必要があろう。

課題 4　MRM 法を使って言語テストの多値 (polytomous) データの規準設定を効果的に行うことは可能か。

課題 5　MRM 分析を使って現実的な等化手続きに基づく規準設定を行うことは可能か。

昨今では Griffin et al. (eds.) (2012) のように、21 世紀型スキルの教育が求められている。また、我が国の大学入試改革の動きの中でも、「教科横断型テスト」といった考え方が検討されており、今後は、複合的な構成概念を持つ、open ended な item/task への対応が迫られるであろう。そのようなテストへの応答は、自ずと 0-1 データから、多値型データへと移行していく。多値型データの形で得られる回答データを処理することも見据えて、規準設定の方法論を検討する必要がある。前出した WINMIRA をはじめ、複数のデータ分析ソフトは、すでに polytomous モデルの処理を可能にしている。そして、その分析に基づき、課題 1〜3 と同様の検討を追求し、現実的な等化手続きに基づく規準設定の方法論について慎重に検証しながら、それぞれの手順のステップについて明快な解説と、ユーザー・フレンドリーなインターフェースを持つ、規準設定のためのプログラムの開発に結びつけていくことが重要である。

以上のように、英語能力の規準設定の方法論に関する研究は緒に着いたばかりであり、課題の解決に向けて継続的な研究を進めていくことが不可欠である

参考文献

American Educational Research Association, American Psychological Association and National

Council On Measurement in Education. (2014). *Standards for educational psychological testing*. American Educational Research Association.

Cummings, K. D. & Petscher, Y. (eds.) (2016). *The fluency construct: Curriculum-based measurement concepts and applications*. NY: Springer.

Downing, A. D. & Haladyna, T. M. (eds.) (2006). *Handbook of test development*. Mahwah, NJ: Laurence Erlbaum associates, Publishers.

Ferrara, S., Hohnson, E. & Chen, W-H. L. (2005). Vertically articulated performance standards: Logic, procedures, and likely classification accuracy. In Cizek, G. J. (ed.) *Applied measurement in education*. Vol. 18, Number 1, 2005: 35-60.

Frick, H., Strobl, C., Leisch, F. & Zeileis, A. (n. d.). Flexible Rasch Mixture Models with Package psychomix. Retrieved from https://cran.r-project.org/web/packages/psychomix/vignettes/raschmix.pdf

Griffin, P., McGaw B. & E. Care. (eds.). (2012). *Assessment and teaching of 21st century skills*. Springer.

Hagenaars, J. A. & McCutcheon, A. L. (eds.) (2002). *Applied latent class analysis*. Cambridge: CUP.

Huynh, H. & Scheneider, C. (2005). Vertically moderated standards: Background, assumptions, and practice. In Cizek, G. J. (ed.) *Applied measurement in education*. Vol. 18, Number 1, 2005: 99-114.

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.

Lane, S., Raymond, M. R. & Haladyna, T. M. (eds.) (2016). *Handbook of test development. (2nd ed.). NY: Routledge.*

Templin J. & Jiao, H. (2012). Applying model-based approaches to identify performance categories. In Cizek, G. J. (Ed.). *Setting performance standards* (*2nd ed., pp.* 379‑397). NY: Taylor & Francis.

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

Rasch Measurement Analysis Software Directory at http://www.rasch.org/software.htm

大友賢二 研究代表・渡部良典 研究副代表. (2012). 『言語テストの規準設定 報告書第1号』 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・渡部良典 研究副代表. (2013). 『言語テストの規準設定 報告書第2号』 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. (2013a). 「予備調査：CITO variation on the bookmark method」. 『言語テストの規準設定 報告書第２号』. 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. 1-38).
大友賢二. (2013b). 「英語教育とテスト：第二言語習得における規準設定をめぐって」. 『第７回日本テスト学会賞記念講演』. 東京：成蹊大学.
大友賢二 研究代表・渡部良典 研究副代表. (2014). 『言語テストの規準設定　報告書第３号』　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・中村洋一・法月　健 (2015). 「英語教育研究センター 2015 年度委託研究中間報告：Mixture Rasch Model による英語能力の規準設定」. 公益財団法人日本英語検定協会.
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」. 『言語テストの規準設定：報告書第３号』. 益財団法人英語検定協会英語教育センター委託研究.

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

３．３．検討結果と今後の課題

大友賢二

本研究に関する２０１５年度委託研究進捗状況報告は、２０１５年１０月１６日に行った。その中で、４．MRM（Mixture Rasch Model）分析の今後の課題として示したものは、５項目であった。課題１　小グループのデータ分析、課題２　MRM 分析後に行う規準設定手続きの条件、課題３　MRM 法と RM-LR 法とのより合理的な比較検証、課題４　多値データーの規準設定、課題５　現実的な等価手続きに基づく規準設定、というものである。これに関連して、研究員の３名にとってそれぞれもっとも重視したい課題を取り上げて、年度末の報告としてその研究成果を纏めることにした。

筆者のとりあげたい課題は、課題２「MRM 分析後に行う規準設定手続きの条件」のなかの、「分割点」を決定する計算手続きに関する検討結果と今後の課題についてである。

１．　これまでの検討結果：

１−１．　データ分析結果：

これを検討してみると、たとえば、５クラスに分類する場合など、それぞれ複数のクラスにおける分割点を求めていることが多い。しかし、受験者全員のクラスを２とした場

合、その分割の決定は可能であるかということをここでは取り上げたい。

その研究結果の一つは、法月(2015)に見いだすことが可能である。その設定は日本人２１３名に対する５０問の英語テスト結果を MRM（Mixture Rasch Model）を用いて分析したものであるが、主観的な判断を用いなくとも、合否判定規準は、下記のように設定することが可能であるということが確認できた。これは重要な発見の一つである。ただし、交点の０．９３３以上は、人数的には、同じ 117 名であるが、この中には、Ｃ１の上位層が一部含まれ、逆にＣ２の下位層が一部除去されている結果となっている点は、今後の課題である。

１−２．　２次方程式の解き方：

MRM を使った規準設定手順を考える場合、欠かせない手法の一つは、２次方程式の解き

方である。これには、３つの方法が考えられている。（A）では、x の１次の項がないもので、x^2=6　=> +/-SQR(6)のように変形するものである。（B）は、x の１次の項があるもので、

x^2+5x+6 =0　=> (x+2)(x+3) =0　−> x+2=0 または、x+3 = 0 −> x = - 2 または、x = - 3 のように、因数分解で解くものである。（C）は、x の１次の項があるもので、因数分解ができないものは＜解の公式＞で解くというものである。ここでは、ax^2+bx+c=0 −>x＝（−b+/-SQR(b^2 – 4ac)）/2a　が考えられている。共同研究者３人がここで用いる手順では、この３つのうち（C）の方法である。

法月（2015）の分析結果を基盤として、様々な検討を重ねた結果、以下、1-3, 1-4 のような結論に達することができた。

１−３．　２つのクラスに分析した場合のデータ：

合否を示す２つのクラスの受験者数、割合、平均、標準偏差は、以下のとおりである。

	N	%	Mean	SD
C1:	96	44.9	-0.044	0.628
C2:	117	54.7	2.202	0.857

１−４．　交点の計算手順に関するデータ：

$$\ln\left(\frac{W_2\sigma_1}{W_1\sigma_2}\right) = \frac{(x-\mu_2)^2}{2{\sigma_2}^2} - \frac{(x-\mu_1)^2}{2{\sigma_1}^2}$$

W_1: 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$_{(1)}$に所属する受験者数の全体における割合 = 0.449

W_2: あるクラス$_{(1)}$に隣接するより平均値のより高いクラス$_{(2)}$に所属する受験者数の全体に

おける割合 = 0.547

σ_1:平均値がより低いクラス$_{(1)}$の得点の標準偏差 = 0.628

σ_2:平均値がより高いクラス$_{(2)}$の得点の標準偏差 = 0.857

μ_1:クラス$_{(1)}$の得点の平均 = − 0.044

μ_2:クラス$_{(2)}$の得点の平均 = 2.202

① 左辺の計算 → = ln ((0.547*0.628)/(0.449*0.857)) = −0.113

② 右辺の 2×(σ_2^2)の計算 → = 2*(0.857 ^2)=1.46 9

③ 右辺の 2×（σ_1^2）の計算 → = 0.78 9

④ 右辺の分母を消す計算（②×③） → = 1.469*0.789=1.159

⑤ ④の際の左辺の計算（①×④） →= −0.113*1.115 = −0.131

①〜⑤の計算プロセスを経て、$ax^2+bx+c=0$ を導くことができ、x の計算が可能となる。（2 次方程式の解の公式）

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

⑥ a = ③−② → = 0.789 − 1.469 = − 0.680

⑦ b = $-\mu_2$（平均点が高い方のクラスの平均）×2×③−($-\mu_1$)（平均点が低い方のクラスの平均）×2×② → −2.202*2*0.789−(−（−0.044）)*2*1.469＝−3.604

⑧ c = μ_2^2×③−μ_1^2 ×②−⑤ → = 2.202^2*0.789−（−0.044）^2*1.469−（−0.131） = 3.954

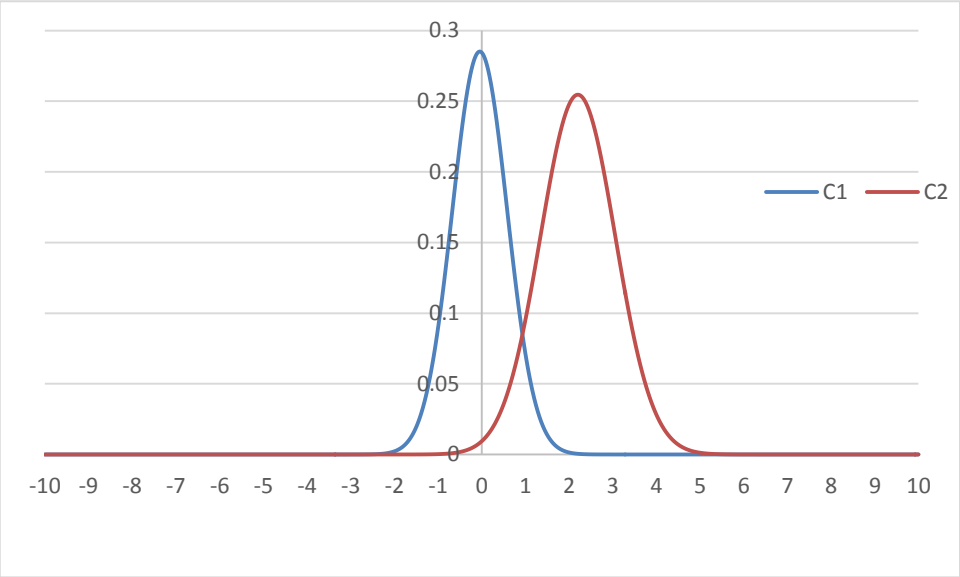
⑨ この２次方程式を解くと、C1 と C2 の平均値の間に位置することが多い値(x_1)と C1 正規曲線と C2 正規曲線の左側の裾で交わる地点($x2$)の値が算出されるが、$x1$ を規準設定の分割点と見なすことができる。

$x1$ = （− （−3.604）−SQR（（−3.604）^2 − 4*(−0.680)*3.954 ））/(2*(−0.680)) ＝ 0.933

x2 = （− （−3.604）＋SQR（（−3.604）^2 − 4*(−0.680)*3.954 ））/(2*(−0.680)) ＝ −6.233

クラスの所属数の比率（もしくは素数）を縦軸、横軸に能力値が標示されるように、正規分布曲線の公式と Excel の機能を使って分布の形状を視覚化すると、図 1 のようなグラフになった。C1 の平均値 (-0.04) と C2 の平均値 (2.20) の間に交点 x_1(0.933) が位置していることが示されている。法月（2015）

図１ Ｃ１とＣ２の正規分布曲線 (VKS データ)



⑩　以上の算出結果をもとにして、ax^2+bx+c=0 の関係が成立するかどうかを確かめると、以下のように成立することが確認でき、分析結果の正確さが確認可能である。

x1=0.933 の場合：

ax^2+bx+c = 0　-0.680*(0.933^2) + (-3.604)*0.933 +3.954 = - 0.592 - 3.363 +3.954 = - 0.001

x2= - 6.233 の場合

ax^2+bx+c = 0　-0.680*((-6.233)^2) + (-3.604)*(-6.233) + 3.954 = -26.418+22.464 +3.954 = 0.000

1－5．　交点設定手順の再確認：

上記の WINMIRA の Mixture Rasch Model を基にした分析手順は、はたして正しかったのかどうかを検討するために、Jiao et al (2011:521) で示されているデータで検討することとした。その結果、以下のように計算手順は正しかったことを確認することが可能であった。

Jiao et al (2011:521) では、W１＝0.059, W2=0.132, σ１＝0.330, σ２＝0.297, u1= -2.460, u2= -1,200 と、上記の２つの式のうちの１つだけが示されている。そして、x1 = -1.876, x2= 10.210 が示されているが、どのような計算手順を経て、x１、x２が求められているのかは、全く示されていない。

そこで、１－４ に示されている計算手順に示されている説明にそって、この分割点 x１、x２ はどのようにして求められたかを示すと、以下の通りとなった。

① 左辺の計算＝＞ln ((0.132*0.330)/(0.059*0. 2 97)) = 0.911

② 右辺の 2 σ 2^2 ＝＞2*(0.297^2) = 0.176

③ 右辺の 2 σ 1^2 ＝＞2*(0.330^2) = 0.218

④ 右辺の分母を消す計算(②*③)＝＞0.176*0.218 = 0.038

⑤ ④の際の左辺の計算（①*④）＝＞0.911*0.038 = 0.035

⑥ a = ③ - ② ＝＞0.218 – 0.176 = 0.042

⑦ b = - u2*2*③ - (- u1)*2*② ＝＞- (- 1.200) ***2***0.218 – (- (- 2.460))

*2*0.176 = - 0.343

⑧ c = u2^2*③ - u1^2 *② - ⑤ ＝＞

（- 1.200）^2*0.218 – (-2.460)^2 *0.176 – 0.035 = - 0.786

⑨ x1 = (- (- 0.343) – SQR ((- 0.343)^2 - 4 * 0.042 * (- 0.786))) / (2*0.042) = - 1.865

x2 = (-(-0.343) + SQR ((- 0.343)^2 -4*0.042*(- 0.786))) /(2*0.042) = 10.032

Jiao et al (2011: 521) で示されているデータは、x1 = - 1.876, x2 = 10.210 であるが、その基になっているデータは小数点以下も含めての計算である。ここで用いた数値は、小数点第３位までのデータを使っているものであり、わずかの相違は問題ではない。従って、Jiao et al (2011)の結果は、上記（３）で示した分割点設定のための計算手順と同じと考え、１－４ で行った計算手順は適切であると判断することができる。

⑩ 以上の算出結果をもとにして、ax^2+bx+c=0 の関係が成立するかどうかを確かめると、以下のように成立することが確認でき、分析結果の正確さが確認可能である。

x1= -1.865 の場合：

ax^2+bx+c =0 0.042*((-1.865)^２) + (-0.343)*(-1.865) + (-0.786) = 0.146 + 0.639 - 0.786 = - 0.001

x2 = 10.032 の場合：

ax^2+bx+c =0 0.042*(10.032^2) + (-0.343)*(10.032) + (-0.786) = 4.227 – 3.441 – 0.786 = 0.000

２． 今後の課題について：

２－１． **A man with two watches is never sure.**

われわれの研究課題である Standard setting （規準設定）や cutting scores （分割点）の行方に関して、筆者の心のなかにいつも残っているものの一つは、North (2014: 222)に記されている以下のような言葉である。

> In contrast to these sensible approaches, Cizek and Bunch (2007), the current US textbook on
>
> standard - setting, explicitly advise *against* using two methods, precisely because these might yield different results. They state that ‘ a man with two watches is never sure’ and ‘use of multiple methods is ill advised’ (Cizek and Bunch 2007:319-20) . Yet replication is the basis of Western academic thought: if you cannot replicate a result you donot have a result!

つまり、規準設定に関しては、２個の時計・方法では、どちらが正しい時間・方法であるかがわからないので、数個の時計・方法を用いるのは適切ではない、と Cizek and Bunch (2007) では述べている。それに対し、replication こそは西洋の学術的考察の基盤であると North (2014)が反論していることである。つまり、誤りを明らかにするための反復実験の必要性を述べている。承知のとおり、North の発言は、EnglishProfile Studies: The CEFR in Practice でのなかのもので、彼は CEFR 研究では、その第一人者でもある。また、Cizek は、National Council on Measurement in Education (NCME) の President の経験もあり、Cizek (Ed.). (2012). *Setting Performance Standards: Foundations, Methods, and Innovations* (Second Edition) などでも、その活躍は広く認められている。

この２つの流れを捉えながら、そして、さらに、Cizek and Earnest (2016: 212-237) で述べられている規準設定や分割点のその後の究明をさらに進めなければならない。

２−２．　CUT SCORES in STANDARDS (2014)

AERA, APA, and NCME (2014). *Standards for Educational and Psychological Testing* は、教育測定の分野での最も注目されている文献の一つであるが、その最新版は 2014 年に顔を出している。そのなかには、第５章として、Scores, Scales, Norms, Score Linking, and Cut Scores が論ぜられているが、そのなかの規準設定や分割点の動向についてもつねに注目しておく必要がある。特に AERA, APA and NCME (2014: 100)で示されている以下の言及は、注目に値する。

> These examples differ in important respects, but all involve delineating categories of examinees on the basis of test scores. Such cut scores provide the basis for using and interpreting test results. Thus , in some situations, the validity of test score interpretations may hinge on the cut scores. There can be no single method for determining cut scores for all tests or for all purposes, nor can there be any single set of procedures for establishing their defensibility. In addition, although cut scores are helpful for informing selection, placement, and other classifications, it should be acknowledged that such category decisions are rarely made on the basis of test performance alone.

つまり、すべての目的のため、あるいはすべてのテストのための分割点を設定する一つの方法は存在しないということである。また、それが正当と認められることを確立するための唯一の手順設定というものも存在するものではない、ということである。

しかし、このことは、規準設定や分割点は設定することに意味がないということではないと判断する。該当するその場において、最も適切な規準設定・分割点設定を究明しな

ければならないということであろう。しかし、それは、すべての分割点設定に、すべての規準設定に適応する唯一の方法はないということに注目しなければならない。つまり、関係者が直面するその場の最善の策を求めなければならないということを意味するものと考える。このことは、科学に関する意思決定のための一般的な基盤でもあろうと考えられる。

２−３．　TOEFL, CEFR and CSE

最近の英語教育に頻繁に顔を出す「can-do リスト」と関連して、CEFR (Common European Framework of Reference) との関連性が論じられている。英語能力の達成目標として can-do リストの設定が取り上げられているが、CEFR との関係は、どうなっているかというものである。多くの研究結果がみられるが、しかし、その関連性をより科学的に究明する方法は、十分検討されているとは言えない。

米国の ETS は、TOEFL (Test of English as a Foreign Language)の開発で有名であるが、そこでの TOEFL iBT Research Report (TOEFLiBT-06) (2008) および Research Memorandum : ETS RM-15-06 (2015)での研究は注目に値する。前者の研究論文は、Tannenbaum & Wylie (2008), 後者の研究論文は、Papageorgiou, Tannenbaum, Bridgeman & Cho (2015：*i*) で示されているが、ここでは、TOEFL と　CEFR との得点の関連性について注目すべき研究結果が見られる。

Based on the feedback of subsequent users and decisions makers, ETS revised the CEFR cut score (i.e. minimum test scores required for each CEFR level) in 2014. In this research memorandum, we present the rationale for the revision of the CEFR cut scores and offer validity evidence that the revised cut scores (a) are reasonable and (b) do not negatively impact the quality of admissions decesions.

Papageorgiou, Tannenbaum, Bridgeman & Cho (2015:6) では、次のように示されていることにも注目する必要がある。前者の論文：Tannenbaum and Wylie (2008) で用いられた規準設提法は、一つには選択肢形式項目に対する modified Angoff approach （Cizek & Bunch 2007 参照）であり、もう一つは、constructed –response items に対する　performance profile approach（Hambleton, Jaeger, Plake & Mills 2000）である。調査のためには、１６の国から参加した２３名の試験官に協力を依頼した結果が示されている。こうした方法、及び、分割点の設定法に関しては、さらなる調査が必要である。

また、英検の CSE スコアー（ Common Scale for English. 2.0 ）では、CEFR, CSE2.0, IELTS, TEAP, 英検１級、準一級、２級　（４技能に対応したスコアー）に関して、例えば、１級（満点３４００、合格２６３０）、準１級（満点３０００、合格２３０４）、２級（満点２６００、合格１９８０）などが示されている。その意味と規準設定の手順に関する考

察は、本研究で行っている規準設定・分割点設定に対しどのように関わってくるのかはさらに究明していくことが必要であろう。

参考文献

AERA, APA, and NCME (2014). *Standards for Educational and Psychological Testing.* AERA.

Cizek, G.J. & Bunch, M.(2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* London , UK: Sage.

Cizek, G.J. (Ed.) (2012). *Setting Performance Standards: Foundations, Methods, and Innovations (Second Edition)*. Routledge .

Cizek, G.J. & Earnest, D.S. (2016). Setting Performance Standards on Tests. In Lane, S. , Raymond, M.R. and Haladyna, T.M. (Eds.). *Handbook of Test Development* (2nd Edition). (pp.212-237). Routledge

Hambleton, R.K., Jaeger, R.M., Plake, B.S., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement*, 24(4), 355-366.

Jiao, H., Lissitz, R.W., Macready, G., Wang, S., & Liang, S. (2011). Exploring levels of performance using mixture Rasch model fo standard setting. *Psychological Test and Assessment Modeling,* 53, 499-522.

North, B.(2014). *The CEFR in Practice*.Cambridge, UK: Cambridge University Press.

Papageorgiou, S., Tannenbaum, R.J., Bridgeman, B., and Cho, Y. (2015). The Association Between TOEFL iBT Test Scores and the Common European Framework of Reference (CEFR) Levels. *ETS RM-15-06.*

Tannenbaum, R.J., & Wylie, E.C. (2008). Linking English-Language Test Scores Onto the Common European Framework of Reference: An Application of Standard-Setting Methodology. *TOEFL iBT Reasearch Report.* ETS

法月　健　(2015). norizuki@ssu.ac.jp　c2_vks1_data （分析後データ xlsx ）.

＊＊＊＊＊