公益財団法人　日本英語検定協会

2016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(2)

2017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

2016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(2)

1.　法月　健：　期待と課題

1.1. 正規分布曲線の交点計算する方法

1.2. 問題点と可能性

2.　法月　健：検討結果と今後の課題

2.1. MRM と LRT との比較検証

2.2. 多値データの規準設定

2.3. 検討結果と今後の課題

3.　中村洋一：今後の課題

3.1. CEFR と他のテストとの関連づけにおける
規準設定の方法

3.2. Standard setting とその活用法

4.　大友賢二：　検討結果と今後の課題

4.1.「特別講演会」における発表内容(ノート)

＊＊＊

Mixture Rasch Model による規準設定：期待と課題

法月　健（静岡産業大学教授）

09/25/2016

1．1．正規分布曲線の交点を計算する方法

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。

Jiao, Lissitz, Macready, Wang & Liang (2011) は、10,000 人の読解テスト（シミュレーション）データ 40 項目について、mdltm と呼ばれるソフトウェアを使って受験者を 5 クラス (階層) に分割する MRM 分析を行い、確率密度関数の概念に基づき、下記の数式（1）から 2 次方程式の解の公式を導き、C1～C5 の隣接するクラス間で正規分布曲線が交わる地点を計算している (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_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$_{(I)}$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) $e^2 = 2.71828182845904^2 = 7.389056099$

公式 (1) の x の値を求めるために、以下の $(L1)$～$(L3)$ の自然対数の性質に従って、計算を行う。

$$\ln(ab) = \ln(a) + \ln(b) \ (L1)$$

$$\ln\left(\frac{a}{b}\right) = \ln(a) - \ln(b)\ (L2)$$

$$\ln(e^a) = a\ (L3)$$

公式 (1) の両辺の自然対数をとると以下の式となり、

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

(L1) の性質に従って、

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

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

さらに (L2) の性質を左辺にあてはめると、次のような式に展開することができる。

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

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

(L3) を右辺に、(L1) を左辺にあてはめると、(2) の式に展開することができる。

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

(2) の式を基に、x の値を計算することができる。

$$\frac{x^2 - 2\mu_1 + \mu_1^2}{2\sigma_1^2} - \frac{x^2 - 2\mu_2 + \mu_2^2}{2\sigma_2^2} - \ln\left(\frac{W_1\sigma_2}{W_2\sigma_1}\right) = 0$$

x と x^2 の項をまとめて、

$$\left(\frac{1}{2\sigma_1^2} - \frac{1}{2\sigma_2^2}\right)x^2 - \left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}\right)x + \left(\frac{\mu_1^2}{2\sigma_1^2} - \frac{\mu_2^2}{2\sigma_2^2} - \ln\left(\frac{W_1\sigma_2}{W_2\sigma_1}\right)\right) = 0$$

両辺に $2\sigma_1^2\sigma_2^2$ をかけると (3) の式になる。

$$(\sigma_2^2 - \sigma_1^2)x^2 + (2\mu_2\sigma_1^2 - 2\mu_1\sigma_2^2)x + \left(\mu_1^2\sigma_2^2 - \mu_2^2\sigma_1^2 - 2\sigma_1^2\sigma_2^2 \ln\left(\frac{W_1\sigma_2}{W_2\sigma_1}\right)\right) = 0 \qquad (3)$$

$ax^2 + bx + c = 0$ のとき $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$ が 2 次方程式の公式の解となるため、(3) の式を展開すると (4) の式になる

$$x = \frac{-(2\mu_2\sigma_1^2 - 2\mu_1\sigma_2^2) \pm \sqrt{(2\mu_2\sigma_1^2 - 2\mu_1\sigma_2^2)^2 - 4(\sigma_2^2 - \sigma_1^2)\left(\mu_1^2\sigma_2^2 - \mu_2^2\sigma_1^2 - 2\sigma_1^2\sigma_2^2 \ln\left(\frac{W_1\sigma_2}{W_2\sigma_1}\right)\right)}}{2(\sigma_2^2 - \sigma_1^2)} \quad (4)$$

(4) の式を使って、μ_1 と μ_2 の間に位置する x の値が隣接する 2 つのクラスの交点を示すことができる。

実際に大友他 (2016) が行った VKS データの 2 クラス分析から、下記のクラス 1、クラス 2 の受験者の比率、平均、標準偏差の値を代入すると、

$$W_1 = 0.4485, \quad \mu_1 = -0.043, \quad \sigma_1 = 0.6276$$
$$W_2 = 0.5467, \quad \mu_2 = 2.202, \quad \sigma_2 = 0.8566$$

2 次方程式の係数は以下の値になる。

$$a = \sigma_2^2 - \sigma_1^2 = 0.3399$$
$$b = 2\mu_2\sigma_1^2 - 2\mu_1\sigma_2^2 = 1.799$$
$$c = \mu_1^2\sigma_2^2 - \mu_2^2\sigma_1^2 - 2\sigma_1^2\sigma_2^2 \ln\left(\frac{W_1\sigma_2}{W_2\sigma_1}\right) = -1.975$$

よって、x は以下の 2 つの値をとる。

$$x = \frac{-1.799 \pm \sqrt{1.799^2 - 4\times0.3399\times(-1.975)}}{2\times0.3399}$$
$$x = -6.225,\ 0.9331$$

(4)の式を簡略化した(5)の式からも、同じ結果を得ることができる。

$$x = \frac{(\mu_1\sigma_2^2 - \mu_2\sigma_1^2) \pm \sigma_1\sigma_2\sqrt{(\mu_1 - \mu_2)^2 + 2(\sigma_2^2 - \sigma_1^2)\left(\ln\left(\frac{W_1\sigma_2}{W_2\sigma_1}\right)\right)}}{\sigma_2^2 - \sigma_1^2} \quad (5)$$

Jiao et al. (2011) の解釈に従うと、μ_1 と μ_2 の間に位置し、クラス 1 のクラス 2 の正規分布分布曲線の交点を示す x の値 0.9331 を、クラス 1 とクラス 2 の分割点と見なすことができる。

大友他 (2016) は、Jiao et al. (2011) に基づき、213 人の語彙テストデータ 40 項目について、WINMIRA 2001 (von Davier, 2001) を使って受験者を 2〜5 クラスに分割する MRM 分析を行い、その結果を報告している。図 1 （原典の図 2、 p.13）は 5 クラス分析の C2 と C3 の正規曲線の交差状況、図 2（原典の図 1、 p.12）は C3 と C4 の正規曲線の交差状況を示している。

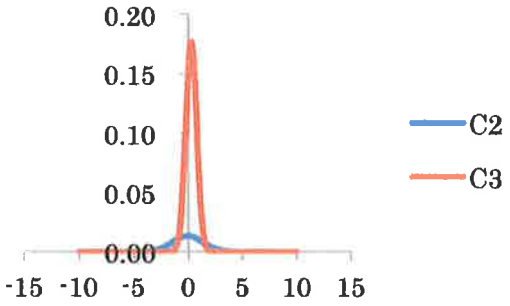


図 1　5 クラス分析の C2 と C3 の正規曲線

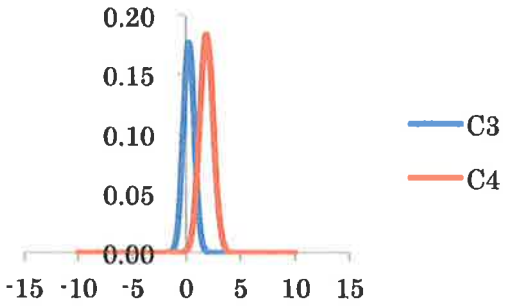


図 2　5 クラス分析の C3 と C4 の正規曲線

大友他 (2016) の語彙テストデータにおける分析では、2 クラス、3 クラスの分析においては、いずれのクラス間においても分割点を設定することができたが、4 クラス、5 クラスの分析では、図 1 のように、クラス間に適切な分割点を得られないこともあった。

1．2．問題点と可能性

Jiao et al. (2011) の研究はシミュレーションデータを使って、理想的な規準設定の手順を示したものであり、正確な分析ができるように、大きなデータ数が確保され、各クラスの平均とクラス間の交点の位置関係が明確に仮定されている。Templin & Jiao (2012) が指摘するように、この規準設定法を様々な現実のデータに応用し、現実の状況下でどのような要因によって誤ったクラス分類が生じてしまうのか、今後さらに注意深く調べていく必要があるだろう。

Kreiner, Hansen & Hansen (2006) と Kreiner (2007) は MRM と関連の統計手法を使って複合的な視点から分割点設定を行っているが、MRM 指標については、WINMIRA 2001 の結果データから比較的な簡易な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 各クラスに所属する受験者の割合を示す混合比率 (mixing proportion) の対比から分割点を探る方法を提案しており、MRM を適用した分析法の一つとして注目される。

大友他 (2016) は語彙テストを使って、Kreiner (2007) と Kreiner et al. (2006) の MRM 分析手続きも試行して、2 クラスを分割する地点を探っているが、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。Jiao et al. (2011) の方法に従って 2 クラスの交点を計算した際には、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し直すと、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、一般のテスト利用者にとって解釈しやすく、利用しやすい情報だと思われるが、ラッシュモデルと異なり、MRM の能力パラメタ—値は素点と一対の関係にはない。実際、Jiao et al. (2011) に基づく 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 点にほぼ相当する。一方で、このパラメタ—値は、C1 の素点では 32 点と 33 点の間に位置すると考えられ、状況の解釈は容易ではない。規準設定の様々な状況における MRM の能力パラメタ—値から素点への換算の可否については、さらに検証を続ける必要があるだろう。

MRM の分析から得られる潜在クラスの編成は、能力以外の要因によっても顕著な影響を受けることがあるように思われ、そのような影響の特徴を十分に理解したうえで、どのような手順で MRM に基づく規準設定を行うことが可能なのか否か、さらに追究していく必要がある。そのためには、MRM 以外の様々な統計的手法を併用し、結果について公正な比較検証を続け、一般のテスト利用者にとっても明瞭な結果が提供され、現実の制約下においても、適切に実践することが可能な、より客観的な規準設定の手続きを構築していくことが求められる。

参考文献

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.

von Davier, M. (2001). WINMIRA2001 [Computer software]. Groningen, the Netherlands; ASC Assessment Systems Cooperation. USA and Science Plus Group.

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.

大友賢二・中村洋一・法月健 (2016).「英語教育研究センター2015年度委託研究報告: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

法月　健（静岡産業大学教授）

02/05/2017

2．1．MRM と LRT との比較検証

大友・中村・法月 2016) において、法月は、日本の大学生 213 名に実施された 50 問の語彙テストについて、WINMIRA 2001 (von Davier, 2001) を使って2～5クラス MRM 分析を行い、Jiao, H., Lissitz, B., Macready, G., Wang, S., & Liang, S. (2011) が提案した下記の数式(1)に基づき、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ている (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_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$_{(I)}$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) $e^2 = 2.71828182845904^2 = 7.389056099$

表1は、WINMIRA 2001 の初期設定の状態で分析を行った後、表計算ソフトウェア Excel を使ってクラス別に関連する統計値を計算し、その結果をまとめたものである。当初、所属受験者数が多い潜在クラス(グループ)から順に、クラス1、2、...とクラス番号が付与されるが、今回は能力による規準設定の目的で MRM 分析を行うため、各クラスの能力推定平均値が昇順になるように、クラスの順番を並べ替えている。よって、能力推定平均値が隣接するCxとCx+1 の2クラス間において、上記の数式の左辺と右辺の x の値が一致して、両クラスの平均

値の間に位置する場合にのみ、示された交点を妥当な分割点と見なした。2、3クラス分析では、すべての隣接するクラスにおいて、平均値の間に位置する妥当な分割点を設定できたが、4、5クラス分析では、分割点が設定できない状況が確認された。たとえば、5クラス分析の C1 と C2 間の交点は-0.89 であったが、この値は C1 の平均値 -0.78 よりも低く、両クラスの平均値間に位置していないため、2つのクラスを分割する妥当な地点ではないと言える。

表1 MRM(2～5クラス)分析による VKS テストデータの分割点設定の可否

5クラス分析							4クラス分析						
	N	$\bar{x}$	SD	ｸﾗｽ間	交点	分割		N	$\bar{x}$	SD	ｸﾗｽ間	交点	分割
C1	16	-0.78	0.63	C1/C2	-0.89	×	C1	12	-0.59	1.36	C1/C2	-1.41	×
C2	48	-0.07	0.50	C2/C3	1.10	×	C2	64	-0.13	0.54	C2/C3	0.42	○
C3	10	0.33	1.38	C3/C4	-0.58	×	C3	38	0.68	0.46	C3/C4	0.91	○
C4	61	0.74	0.54	C4/C5	1.25	○	C4	99	1.71	0.72			
C5	78	1.89	0.68										
3クラス分析							2クラス分析						
	N	$\bar{x}$	SD	ｸﾗｽ間	交点	分割		N	$\bar{x}$	SD	ｸﾗｽ間	交点	分割
C1	47	-0.49	0.57	C1/C2	-0.13	○	C1	96	-0.10	0.67	C1/C2	0.69	○
C2	85	0.61	0.63	C2/C3	1.26	○	C2	117	1.62	0.72			
C3	81	1.87	0.69										

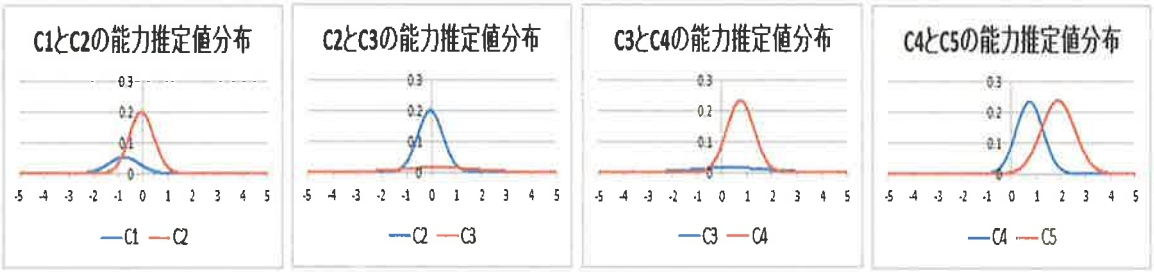


図1 VKS テストにおける MRM 5クラス分析の分割点設定 (smooth score distribution)

図 1 は、5 クラス分析の交点の位置を図示したものである。上記の数式 (1)の左辺と右辺の x が同一の値を示す地点を探るために、Excel のワークシート上で-10～+10 の区間を 0.01 ずつ区切って数値を対比する一覧表を作成し、その値を散布図に代入することで、隣接するクラス間の確率密度曲線を表すことができる。右端の C4 と C5 の関係を除き、分割点設定が有効に機能していない状況が示されていると言える。

Kreiner (2007, p.140) の議論を参考にして、さらに、初期設定「Smooth score distribution」のチェックマークを外して、同様の2～5クラス分析を行ったが、初期設定の分析と比べると、総じてクラス間における所属者数の差が大きくなり、規準設定の難しさが示唆される状況が確認された。たとえば、3クラス分析では、C1 が 110 名、C2 が 99 名、C3 がわずか 4 名になる等、C2／C3 において、実質的に有効な分割点が設定できない結果となった。

なお、表1に示した今回の MRM 2クラス分析 (smooth score distribution の条件) の交点 0.69 は、実測値で比較すると、0.58 (素点 29 点) と 0.70 (素点 30 点) の間に位置するが、期待値の場合は、クラスによって値がやや異なる。たとえば、上記の交点 0.69 は実測値に基づく計算によるものだが、仮に期待値に基づく計算で、交点が同じような値を示した場合、平均値が相対的に低いクラス C1 に所属する受験者にとっては、分割点は素点 29 点と 30 点の間に位置するが、平均値が相対的に高い C2 に所属する受験者にとっては、分割点は 27 点と 28 点の間に位置することになる。一方、各クラスの受験者の MRM 能力実測値を素点の平均値と標準偏差に換算して、交点を求めると、29.96 点という分割点が得られる。今回の2クラス分析(smooth score distribution)においては、いずれの計算でも類似の結果が確認されたが、異なる状況下では、MRM 実測値、MRM 期待値、MRM 実測値から換算した素点のどのスコアを参照するかによって、かなり異なる分割点が導かれることも確認されたため、安定した解釈の確立に向けて、さらなる研究が必要である。

以上の分析は MRM 分析で得られた統計値を基盤にして、分割点となる確率密度関数の交点の算出したものであるが、法月(2014) が VKS データに対して行ったラッシュモデル(RM) と潜在ランク理論 (LRT) を融合して分割点を探る RM-LRT 法に、この方法を適用して、同データの2～5ランク分析を行い、MRM 分析の結果と比較することとした。

まず VKS データのラッシュモデルによる受験者能力推定値を、WINSTEPS (Linacre, 2014) と呼ばれる統計ソフトウェアによって算出し、潜在ランク理論の分析ソフトウェアである Exametrika (荘島、2011) によって各受験者にランクを付与し、上述の MRM 分析と同様の方法で、隣接するランク・グループの確率密度関数の交点を計算した。

VKS データは受験者数が多くないため、Exametrika の分析では、サンプルサイズが小さくても分析することが可能な、個人学習型モデルの自己組織化マップ(LRT-SOM) による推定を行い、できるだけ各潜在ランク・グループに受験者が均等に分布するように、「目標潜在ランク分布」は初期設定の「一様分布」を選択した。表2は、このような手順に従って、分割点設定を行った結果をまとめたものである。図2は、RM-LRT 法の 5 ランク分析において、いずれのランク間でも有効な分割点設定を行うことが可能であることを示している。

表2 RM-LRT 法 (2～5ランク) 分析による VKS テストデータの分割点設定の可否

5ランク分析							4ランク分析						
	N	$\bar{x}$	SD	ランク間	交点	分割		N	$\bar{x}$	SD	ランク間	交点	分割
R1	38	-1.62	0.47	R1/R2	-1.17	○	R1	52	-1.46	0.49	R1/R2	-0.88	○
R2	41	-0.73	0.33	R2/R3	-0.40	○	R2	51	-0.41	0.31	R2/R3	-0.01	○
R3	47	-0.03	0.33	R3/R4	0.32	○	R3	59	0.42	0.33	R3/R4	0.91	○
R4	46	0.65	0.29	R4/R5	1.10	○	R4	51	1.41	0.67			
R5	41	1.53	0.69										
3ランク分析							2ランク分析						
	N	$\bar{x}$	SD	ランク間	交点	分割		N	$\bar{x}$	SD	ランク間	交点	分割

R1	67	-1.29	0.54	R1/R2	-0.61	○	R1	103	-0.94	0.67	R1/R2	-0.06	○
R2	74	0.00	0.39	R2/R3	0.56	○	R2	110	0.88	0.71			
R3	72	1.20	0.67										

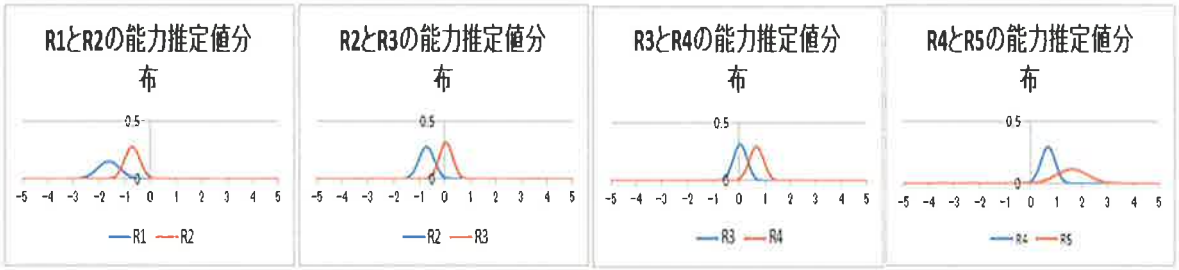


図2　VKS テストにおける RM-LRT 5 ランク分析の分割点設定（一様分析）

以上の結果から、MRM 分析と異なり、RM-LRT 法で分析したすべてのランク間においては、有効な分割点が得られたことが確認できる。MRM と RM-LRT の分析結果をさらに細かく比較すると、前者の分析では、平均値順にグループ（クラス）を並べ替えた後でも、隣接するクラス間で能力値が全体的に近似する場合が少なくなかったが、後者においては、常に明確な能力層のランク集団に分かれていた。前者については、妥当な分割点が得られるクラス間であっても、クラス内で能力値の分散度がやや高く、隣接クラス間で確率密度曲線の重なる部分が大きいこともあったが、後者は、ランク内の能力値の差が相対的に小さく、ランク間で能力値が重なる部分が小さい傾向が見られた。サンプルの受験者数がそれほど多くないことも要因として考えられるが、少なくとも VKS テストデータにおいては、Jiao et al. (2011) が提唱する分割点設定法は、MRM 分析よりも RM-LRT 法に適しているようである。

2．2．多値データの規準設定

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)。この理解度自己評価のデータを利用して、多値データにおける MRM と RM-LRT 法の分割点設定を比較検証することとした。分析の便宜上、「5」、「4」、「2」、「1」の 4 段階尺度は、「3」、「2」、「1」、「0」に変換し、無回答項目が含まれるデータを除く 193 名のデータを分析した。多値データの MRM 分析にも WINMIRA 2001 (von Davier, 2001) を使用したが、各受験者が同一数(4 段階)、同一条件と仮定される評定尺度に回答しているため、初期設定の「Ordinal (Partial Credit) Model」のチェックマークを「Rating Scale Model」に変更した。そして、表1の正誤2値データ分析の時と同

様に、初期設定の「Smooth score distribution」にチェックされている状態と、チェックを外した状態で、2～5クラス分析を行った。表3は2～5クラスの MRM 分析の結果をまとめたものであるが、VKS テスト正誤2値データの分析結果とは異なり、「Smooth score distribution」にチェックがついた初期設定の状態のほうが、クラス間での受験者分布の偏りがより顕著で、4 クラス分析のように、所属者数が 1 名しかいない分析不能なクラスが含まれるケースもあったため、チェックを外した状態で分析を行った結果を提示することとした。

とは言え、「smooth score distribution」でない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、そのうちの大半は交点すら存在しない状況となっている。また、この方法であっても、クラス間で構成人数に相当な差が存在するため、単純なクラス間での能力比較が困難な場合もある。いずれにしても、隣接クラスの正規曲線の形状から見ても、能力以外の質的に識別できる何等かの要因が、クラス形成に強く影響している可能性が高いと言える。図3は、VKS 理解度データの MRM 5 クラス分析において、分割点設定が、いずれの隣接クラス間においても機能していない状況を示している。

表3 MRM(2～5クラス)分析による VKS 理解度自己評価の分割点設定の可否

5クラス分析							4クラス分析						
	N	$\bar{x}$	SD	クラス間	交点	分割		N	$\bar{x}$	SD	クラス間	交点	分割
C1	16	-0.57	2.08	C1/C2	なし	×	C1	12	-0.35	0.56	C1/C2	なし	×
C2	5	-0.53	0.72	C2/C3	なし	×	C2	83	0.85	1.38	C2/C3	-0.30	×
C3	45	0.19	1.00	C3/C4	0.90	×	C3	95	1.06	1.32	C3/C4	5.23	×
C4	37	0.65	1.03	C4/C5	なし	×	C4	3	1.41	4.06			
C5	90	1.68	1.49										
3クラス分析							2クラス分析						
	N	$\bar{x}$	SD	クラス間	交点	分割		N	$\bar{x}$	SD	クラス間	交点	分割
C1	12	-0.35	0.63	C1/C2	なし	×	C1	41	0.07	0.91	C1/C2	なし	×
C2	84	0.88	1.38	C2/C3	-0.70	×	C2	152	1.01	1.32			
C3	97	1.01	1.32										

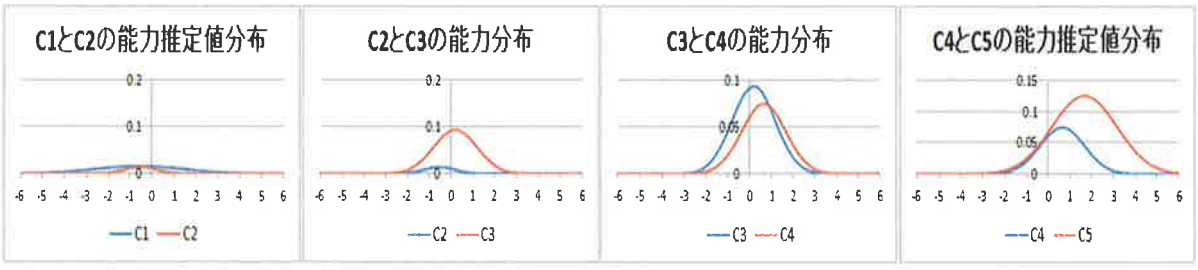


図3 VKS 理解度自己評価における 5 クラス分析の分割点設定 (4 段階評定尺度)

一方、表4のように、MRM 分析結果とは対照的に、RM-LRT 法の2～5クラスの分析においては、すべてのクラス間で有効な分割点が設定できることが確認できた。受験者数が少ない

ため、VKS テストの2値データの分析時と同様に、一括学習型モデルの生成トポグラフィックマッピング(LRT-GTM) による推定はできなかったが、個人学習型モデルの自己組織化マップ(LRT-SOM) による推定は、VKS 理解度自己評価の多値データの分割点設定にも有効であったと言って問題ないであろう(LRT-GTM と LRT-SOM の機能については、荘島 (2010、pp.92-98)を参照)。

表4 RM-LRT 法 (2～5ランク) 分析による VKS 理解度自己評価の分割点設定の可否

5ランク分析							4ランク分析						
	N	$\bar{x}$	SD	ランク間	交点	分割		N	$\bar{x}$	SD	ランク間	交点	分割
R1	34	-1.47	0.53	R1/R2	-1.07	○	R1	44	-1.37	0.50	R1/R2	-0.83	○
R2	41	-0.63	0.31	R2/R3	-0.35	○	R2	49	-0.44	0.23	R2/R3	-0.05	○
R3	42	-0.05	0.27	R3/R4	0.31	○	R3	54	0.36	0.24	R3/R4	0.80	○
R4	40	0.61	0.19	R4/R5	0.98	○	R4	46	1.48	0.86			
R5	36	1.66	0.89										
3ランク分析							2ランク分析						
	N	$\bar{x}$	SD	ランク間	交点	分割		N	$\bar{x}$	SD	ランク間	交点	分割
R1	62	-1.13	0.57	R1/R2	-0.56	○	R1	96	-0.85	0.61	R1/R2	-0.02	○
R2	68	-0.03	0.37	R2/R3	0.55	○	R2	97	0.90	0.82			
R3	63	1.24	0.83										

2．3．検討課題と今後の課題

VKS テストの正誤2値データの MRM 分析に関しては、5、4クラス分析ではクラス間で有効な分割点設定ができないケースが確認されたが、いずれも所属受験者数が 20 名以下の小さなクラスが含まれる状況であった。今後、分布ができるだけ均等化される分析方法の導入や、Kreiner (2007) や Kreiner, Hansen & Hansen (2006) のような異なる分割点設定法 (大友・中村・法月、2016, pp.28-29) の応用、複合型技能テストのような性格の異なるテストデータを扱った分析の実施等を講じる余地もあるが、単純にもう少し大きな受験者サンプルを使って、分析の有効性を追究する必要があるだろう。

VKS 理解度自己評価の多値データの MRM 分析に関しては、分割点設定の実践化に向けて議論を継続することは困難だと言わざるを得ない。しかしながら、大規模データの検証の可能性を探るとともに、Baghaei & Carstensen (2013) が 2 値データで議論したような、MRM のクラス間で特定の項目難易度が変動する特異項目機能 (differential item functioning) が、VKS 理解度多値データの分析にどのような影響を及ぼしているか、4 段階尺度の 3/2、2/1、1/0 の3つのしきい値水準 (threshold levels) 様式から探ることが期待される受験者の理解度回答を巡る心理作用等、今後も異なる角度から、規準設定に影響を及ぼす要因について、議論を続けていく意義は十分にあると言えるだろう。

一方、RM-LRT 分析においては、2値、多値データともに計測したすべてのクラスにおいて

有効な分割点設定を行うことができたが、サンプルサイズが小さく LRT-GTM による潜在ランク理論の推定を実施することはできなかった。通常の SOM による推定は GTM によるものと異なり、「同じデータでも分析するたびに結果が微小に異なる(荘島、n.d.)」とされているため、受験者の進路に大きな影響を与えるテストの規準設定には、時と状況によっては、LRT-GTM による推定が求められるかもしれない。MRM 分析と同様に、今後はもっと大きな受験者サンプルを対象として、LRT-GTM による推定についても、検証を行う価値があるだろう。

参考文献

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.*

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.

Linacre, M. (2014). *WINSTEPS Rasch measurement computer program* (Version 3.81.0). Chicago: Winsteps.com.

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

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・中村洋一・法月健 (2016).『英語教育研究センター2015年度委託研究報告: 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.

荘島宏二郎 (2010).「ニューラルテスト理論」、植野真臣・荘島宏二郎(編).『学習評価の新潮流』(pp.83-111)、東京:朝倉書店.

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.).「潜在ランク理論」 retrieved from http://antlers.rd.dnc.ac.jp/~shojima/exmk/jindex.htm

Mixture Rasch Model による英語能力の規準設定: 今後の課題

中村 洋一（清泉女学院短期大学教授）

02/04/2017

３．１．CEFR と他のテストとの関連づけにおける規準設定の方法

Manual for relating Language Examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment (CEFR). (以下 Manual, Council Europe, 2009) が提供している CEFR と他のテストを関連づける (aligning) 手順は、大きく分類すると Familiarisation : CEFR の descriptors に関する詳細な知識を身につける準備段階、Specification: Familiarisation の知識に基づき、それと比較する開発中のテストの細目を検討する作業、Standardisation training / benchmarking: CEFR Level と開発中のテストの規準を比較検討し、標準化するトレーニング、Standard Setting: 開発中のテストにおける cut-off scores を設定して規準設定を行い、CEFR との関連づけを決定する、Validation: 関連づけ作業の妥当性を検討する継続的な作業の 5 つの "on track" な手順 (あるいは "stages") である (pp. 7 - 11)。この手順に則って、いくつかの言語テストと CEFR との関連づけが試みられてきている。

Tannenbaum & Wylie (2008) は、TOEFL® iBT, TOEIC®, TOEIC Bridge ™ Test について CEFR levels と関連づけた "minimum scores (cutscores)" (p.1) を明らかにする試みを行い、familiarization と standardization of judgments の手順に焦点をあてて (p. 3) 詳説している。採用した standard setting の方法は、Selected-response sections では modified Angoff approach (pp. 8 – 11)、Constructed-response sections では modified examinee paper selection method (pp. 11 – 12) であった。この試みは "The cutscores were constructed following well-established standard-setting procedures." (p. 28) とまとめ、Table 18 (TOEFL iBT Test), Table 19 (TOEIC Test), Table 20 (TOEIC Bridge Test) に、一覧として提示し (p. 29)、"a significant step forward" (p. 31) であると締めくくっている。

日本における関連づけの検討としては、日本英語検定協会と日本生涯学習総合研究所による先駆的な研究、Common Scale for English: CSE 2.0 開発の取り組み (Brown 他. 2012, Dunlea. 2009, 2010, 2014, Nakamura. 2015) があり、項目応答理論による能力推定値算出処理、垂直的等化・水平的等化の処理といった統計処理を経て、能力推定値をスコア尺度に変換して、ユニバーサルで連続する尺度を算出し、それを基に、英検 Can-do リストと CEFR との比較や、専門家パネルを中心とした Basket 法・Modified Angoff 法による検証を行い、規準設定 (各級の合格点) を決定し、CEFRとの関連づけを行っている。また、大学入学者選抜試験改革における外部テストの活用と関わって進められているいくつかの取り組みは、2016年3月25日開催の平成27年度英語力評価及び入学者選抜に

おける英語の資格・検定試験の活用促進に関する連絡協議会において提出された資料 7 英語の資格・検定試験に関する基礎資料にまとめられている。資料 7-2 は、各種テストの細目について概略的な内容を記し、続いてごく概略の説明に留まるが、「主な英語の資格・検定試験の CEFR との関係性検証方法」の一覧を示している。この一覧は、10種類のテストで規準設定に採用された検証方法をリストアップしており、「Manual にのっとって実施」と記述しているテストが 1 つ、Basket 法・Modified Angoff 法・Yes / No method・Bookmark Methodといった、具体的な規準設定法を提示しているテストが 5 つある。また10種類のテストすべてで、複数の方法を複合的に検証している旨が記されており、CEFR とそれぞれのテスト得点との関連づけにおいて様々な検討が、慎重になされていることを示している。

上記のような試みが継続される中で、Papageorgious et al. (2015) は、2014年に ETS が、上述した Tannenbaum & Wylie (2008) の "panel-recommended cut scores" を再検討し、"lowered those recommendations by 2 SEM" とした報告を公開していることに注目したい。この報告では Tannenbaum & Wylie (2008) で cutscores 設定の判断材料のひとつとした 1 標準誤差信頼区間 (SEM) の 68% を、"the feedback of subsequent users and decision makers (p. i)" に基づいて検討した上で、2 SMEの 98% の範囲へと変更して revised cut scores を Table 1 (p. 8) にまとめ、"the revised cut scores (a) are reasonable and (b) do not negatively impact the quality of admissions decisions (p. i)" と結論を示している。

ひとたび設定した規準を「手直し」することには大きなエネルギーが必要となるであろう。ましてやそれが、資格審査や入学許可のような利害関係が高い判断をも伴うものであれば、社会的責任の観点からみても、変更を呻吟する負の動機になり得るのかもしれない。しかし、Papageorgious et al. (2015) の conclusionにも触れられているとおり、"..., it should not be assumed that the relationship between a language test and the CEFR levels is necessarily simple, direct or established as a one-time event (p. 14)" 、"..., it is appropriate to periodically review and reconsider the relationship between test scores and the CEFR" との指摘を心すべきであり、我が国の研究についても、これまでに取り組まれてきた CEFR との関連づけについて、さらに検討を継続する必要がある。

継続検討の作業は主に Validation と深く関係するものである。Manual では External Validation の方法論が検討されているが (pp.89 - 117)、その議論が "may look disappointing ..., as it does not make a clear distinction between good and bad" とまとめられ、その理由として "there is no authority that owns the truth" と " ... it is not realistic to expect a definite verdict in any particular case." があげられ "It is the responsibility of the present generation to provide the necessary data and documentation for such a meta-analysis to be meaningful (p. 117)" と結ばれている。このある意味「歯切れの悪い」結論は、ひとつの規準を設定すると、それに固執する傾向がなきにしもあらずと見える我が国の言語テストのシステムや教育文化においても真摯に心すべき戒めであり、長期的な視野も鑑みながら、よりよい規準設定の方法を継続的に検討し、適切な対応をしていくことが重要である。

検討の具体的な切り口のひとつとして、「純粋な規準設定」に関する方法論の検討を継続していく必要がある。Manual と併せて、North & Jones (2009) も規準設定方法の継続的検討が不可欠であるとし、Further Material として "the use of scaling techniques" と "item banking" の具体的な方法論に焦点を当てて紹介している。また、*Reference Supplement* (2009) も、規準設定の理論的フレームワークは "still unsolved issues (Foreword, p. 3)" であると警鐘をならしつつ、開発・研究途上の理論的・技術的な解説を提供している。比較的「新しい」とされている IRT による方法論についても、"a caveat to overoptimistic proponents to IRT: using an IRT-model does not convert a bad test into a good one (Section G, p. 15)" と注意を促している。この警告は、規準設定に IRT の枠組みを使用する方法論において看過することは出来ないものである。特に、"If it is qualitative, persons belong to (unobserved) classes or types (of language proficiency); if it is quantitative, persons can be represented by numbers or points on a line. Only the later case is dealt with in Section G. (Foreword, p. 4)" との言及にあるように、少なくても *Reference Supplement* に紹介されている方法論には "qualitative" な視点が含まれていないという指摘に注意を向ける必要がある。より適正な規準設定のためには、本研究の大きな関心事である Mixture Rasch Model の適用により、qualitative な視点を含めて latent group を推定しながらパラメータを計算し、分割点を決定していこうとする方向性の検討が意義深いものであり、規準設定の方法論研究において重要な課題のひとつだと考えるものである。

３．２．Standard Setting とその活用法

前項 3. でとりあげた研究の「前提」を、Standard Setting とその活用法の研究において確認すべき課題のひとつとしてあげたい。Tannenbaum & Wylie (2008) は、その Introduction で、"The study was not intended or designed, however, to establish a concordance between scores on the series of English-language tests, such that scores on one test could be used to identify comparable score on the another test." と宣言し、"Scores from each test were independently mapped to the CEFR levels; no attempt was made to link scores or score distributions across the test. (p. 2)" と明言している。この研究が、あるテストの得点と CEFR の levels との関連づけは、それぞれ独立して行ったもので、複数のテスト得点の間にある関係性を見いだすことに関しては、なんらの検討もしていないと述べている点に注意が必要である。多くの関連づけの研究では Tannenbaum & Wylie (2008) と同様のデザインを採用し、CEFR とそのテストとの関連づけを主な焦点としている。そのような制限がある場合は、CEFR を中心的な規準として、それぞれ複数のテストが独立して関連づけを行った結果を統合し、それを横並びの一覧として検討するというような方法は、統計処理における妥当性を満たしているとは言い難く、また、それぞれのテストの構成概念妥当性が異なるという観点からも、CEFR との関連づけのみを基にした複数のテスト得点間の単純比較は、大変危険なものであると言わざるを得ない。その意味において、2016 年 3 月 25 日に開催された平成 27 年度英語力評価及び入学者選抜における英語の資格・検定試験の活用促進に関する連絡協議会 (以下、連絡協議会) において提出された、資料 7-1「英語の資格・

検定試験に関する基礎資料 各試験団体のデータによる CEFR との対照表 (2016/03/25 版)」にある「※ 各試験団体の公表資料より文部科学省において作成」という脚注に、十分注意する必要があるであろう。

複数のテスト間の関係性を検討するためには、その相関関係を検証し、ひとつのテストのデータ (x) からもう一方のテストの得点 (y) を予測する回帰式、$y = a + bx$ (a=定数項, b= 回帰係数) で表現する試みが行われてきた。日本英語検定協会と日本生涯学習総合研究所による、Common Scale for English: CSE 2.0 開発の取り組みでは、相関関係の分析や、回帰分析と呼ばれる統計的手法を用いて CSE 2.0 のスコアが、他のテストのどのくらいのスコアに該当するのかといった研究もなされ、Brown 他 (2012, pp. 43 - 47) では、次のように CSE から TOEFL iBT の得点を予測している。

Predicted TOEFL iBT score = 86.517 + 13.898 x (total EIKEN CSE)

また、英検の「大学入試センター試験との相関調査 『実用英語技能検定』と『TEAP』で実施」の研究では、3つテストの相関関係を調査し、センター試験 - TEAP : $r = 0.798$, センター試験 - 実用英語技能検定: $r = 0.894$, 実用英語技能検定 – TEAP: $r = 0.844$ といった相関関係を検証している。一般的に 0.800 の相関係数は「高い相関がある」とされているので、この3つのテストは、お互いに相関関係が高く、大学入試判定といった目的としては、いずれかのテストを受験していれば、回帰式を用いた予測により、共通尺度上での比較的正確な解釈が可能になり得るという点で、実用性の追求ができるかもしれない。

しかし一方で、上記連絡協議会に提出された資料 2-2 の、「3. 対照表・換算表について」の「 … それぞれの資格・検定試験は目的や主な受験者層が異なるので国が策定するのは困難と思われる。…、CEFR との対照 (対照表) に関する情報の収集・提供を継続して行うべき」という指摘は、追求すべき課題のひとつである。現在は、構成概念の異なる複数のテストにおいて設定された規準を横並びの一覧表にまとめたものが存在するのみという段階であることを再確認する必要がある。また、検定試験の評価等の在り方に関する調査研究協力者会議 第1回 (平成28年12月6日) 資料 3 (p. 6) においても、「受検者が検定実施団体に求める情報公開の内容」のうち、【検定試験のパンフレットやHPに記載されていると信頼できると思う項目】は、合格基準 (点) の項目で 33.6% と比較的信頼されている度合いが低く、【検定実施団体に情報公開してほしいこと】は、合格基準 (点) の項目で 52.4% と比較的要望が高いとのデータが提示されており、各種テストにおける規準設定方法について、広く、明確に開示し、さらに透明性を向上させていくことも必要不可欠であろう。

今後の研究課題のひとつは、それぞれのテストにおいて設定された規準である複数の cut scores 相互間の関係性をより詳細に検討し、回帰式あるいはそれに代わる関係性を見つけ出し、その妥当性を検証することであると思われる。そして、その過程においては、Standard Setting の結果の活用法を、その目的に照らし合わせて慎重に検討することに留意が必要である。「日本のほとんどの大学入試では、スタンダードは必要ない。それは、日本の大学入試の合否ラインは、定員によって決められているからである。」(根岸 2016) というような発言は、日本の大学入試を取り巻く現状における目的論として捉えるべきであり、言語テストにおける Standard Setting の研究がカバーすべ

き領域のごく一部を批判的に捉えたものにすぎない。昨今の大学入学試験改革の議論の中で拙速が懸念され、ともすれば玉虫色的にも捉えられかねない「換算点」のような概念形成についても、さらにより慎重な検証が必要であろう。

本研究の主な関心事は、設定される規準の活用という目的論の広範な分野と無関係に存在するものではないが、まず、純粋に、言語能力を問う言語テストにおける Standard Setting の方法論を検討することである。Manual には、Standard Setting に関して、Test-centered (TC)・Examine centered (EC)・IRT analysis (IRT) という3 つのカテゴリーに分類された 10 の方法論が詳説されている (pp. 57 - 87)。また、Manual の出版と同じ年に、大友 監修 (2009) が「これまでの分割点設定法」について検討しているが、残念ながら、言語テストにおける規準設定について、当時の日本では大きな議論には至らなかったように思われる。

これからの Standard Setting 研究においては、人間による判断にとどまらず、統計的な見地からより客観性の高い基準を絞り込む方法論を追求することが重要であり、MRM あるいは、LCA 等によるデータ処理手順を検討していくことの有用性、可能性を検討することに大きな意義があると考える。さらには、ライティングやスピーキングのテストにおける段階的な得点・部分点 (partial credit) のスコアを処理していくためには、多値モデル (Polytomous IRT) のためのアルゴリズムや処理過程の選定も必要である。Standard Setting、とりわけ英語能力の規準設定においては、あるひとつの尺度を設定した後も継続的に検討・改善していくことが不可欠であり、新しい知見、新しい統計処理、より正確で客観的な規準設定方法を追求していくことが重要である。

参考文献

Brown, J. D., Davis, J., Takahashi, C., & K. Nakamura. (2012). "Upper-level EOKEN Examinations: Linking, Validating, and Predicting TOEFL®iBT Scores at Advanced Proficiency Eiken Levels". Eiken Foundation of Japan.
(Retrieved from http://www.eiken.or.jp/eiken/group/result/pdf/eiken-toeflibt-report.pdf)

Council of Europe. (2009). *Manual for relating Language Examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment.*
(Retrieved from http://www.coe.int/t/dg4/linguistic/Manuel1_EN.asp, Dec. 2016)

Council of Europe. (2009). *Reference Supplement to the Manual for Relating Language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment.*
(Retrieved from http://www.coe.int/t/dg4/linguistic/Manuel1_EN.asp, Dec. 2016)

Dunlea, J. (2009, 2010). 「英検と CEFR との関連性について 研究プロジェクト報告」. STEP英語情報 2009-11・12月号、2010-1・2月号より転載.
(Retrieved from https://www.eiken.or.jp/center_for_ research/pdf/market/report_02.pdf)

Dunlea, J. (2014). "Incorporating the CEFR into language test development: Using an international framework in local contexts". (PPT file presented in E-merging Forum 4 in Moscow, Retrievef from http://www.britishcouncil.ru/sites/default/files/emerging4_dunlea_public_25032014_pdf_0.pdf)

Martyniuk, W. ed. (2011). *Aligning tests with CEFR: Reflections on using the Council of Europe's draft manual* (Studies in Language Testing 33). Cambridge University Press.

Nakamura, K. (2015). "Investigating the comparability of different level test results using the Rasch model".
(PPT file presented in the 12th EALTA conference 2015, Retrieved from http://www.ealta.eu.org /conference/2015/presentations/Friday/Friday%20room%2023.0.50/Nakamura.pdf

North, B. & Jones, N. (2009). *Relating Language Examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment, Further Material on Maintaining Standards across Languages, Contexts and Administrations by exploiting Teacher Judgment and IRT Scaling.*
(Retrieved from http://www.coe.int/t/dg4/linguistic/Manuel1_EN.asp, Dec. 2016)

Papageorgious, S., R. J. Tannenbaum, B. Bridgeman & Y. Cho. (2015). 'The association between TOEFL iBT® test scores and the Common European Framework of Reference (CEFR) levels.' (Research Memorandum: Retrieved from http://www.toefl.com.tw/pdf/RM-15-06.pdf)

Tannenbaum, R. J. & Wylie, E. C. (2008). 'Linking English-language test scores onto the Common Framework of Reference: An application of standard-setting methodology.' (TOEFL iBT Research Report, ETS: Retrieved from https://www.ets.org/Media/Research/pdf/RR-08-34.pdf)

大友賢二監修 中村洋一・小泉理恵編集. (2009). 『言語テスト: 目標の到達と未到達』. 英語運用能力評価協会.

根岸雅史. (2016). 「初参加のLTRCワークショップは『頭の体操』」. In JLTA News Letter No. 41 (日本言語テスト学会).

日本英語検定協会による文書

「*日本初、英語の4技能テスト結果を比較可能とするユニバーサルなスコア尺度「CSE (Common Scale for English)」現状の研究・開発状況と今後の展望についてのご報告 (H27.3.19)」 (Retrieved from [https://www.eiken.or.jp/association/info/2015/pdf/20150317_pressrelease_cse.pdf]

「大学入試センター試験との相関調査「実用英語技能検定」と「TEAP」で実施 (27-10-07)」 (Retrieved from [https://digitalpr.jp/r/13570]

文部科学省. (2014). 「英語教育の在り方に関する有識者会議 英語力の評価及び入試における外部試験活用に関する小委員会 審議のまとめ」

(Retrieved from http://www.mext.go.jp/b_menu/shingi/chousa/shotou/102/102_2/index.htm)

文部科学省. (2016). 検定試験の評価等の在り方に関する調査研究協力者会議 第1回 (平成28年12月6日)

資料3 検定試験等に関する参考資料

(Retrieved from http://www.mext.go.jp/b_menu/shingi/chousa/shougai/038/index.htm)

文部科学省. (2016). 平成27年度英語力評価及び入学者選抜における英語の資格・検定試験の活用促進に関する連絡協議会 第2回 (平成28年3月25日)

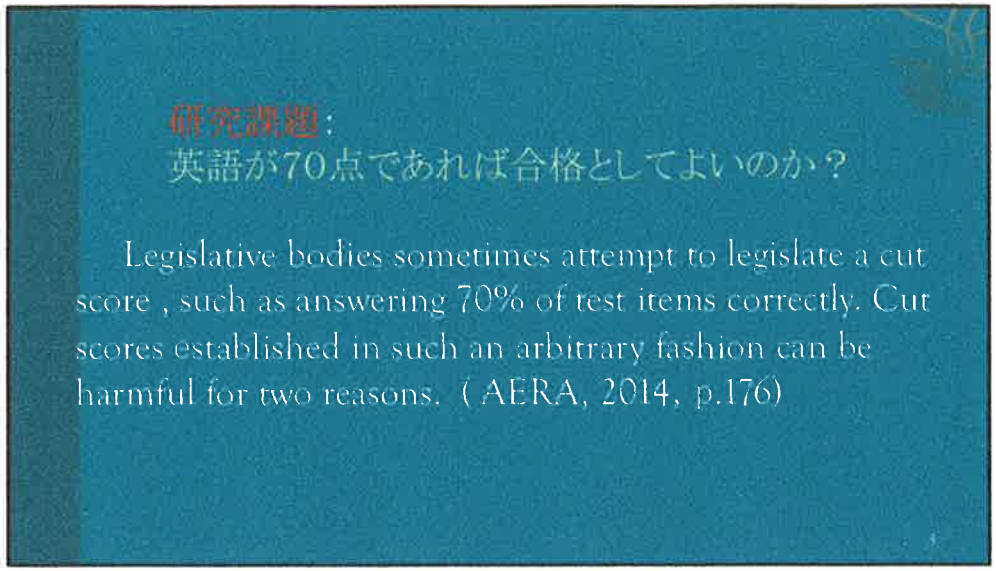
資料 2-2 作業部会における主な意見

資料 7-1 英語の資格・検定試験に関する基礎資料 (各試験団体のデータによる CEFR との対照表 2016/03/25 版)

(Retrieved from http://www.mext.go.jp/b_menu/shingi/chousa/shotou/117/index.htm)


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
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)
日本英語検定協会英語教育研究センター特別講演会 3/18/2017
1

皆さん、こんにちは！！！大友賢二と申します。20分ぐらいのごく短い時間で、「Mixture Rasch Model による英語能力の規準設定」、ということについて、お話しいたします。



＜研究課題＞

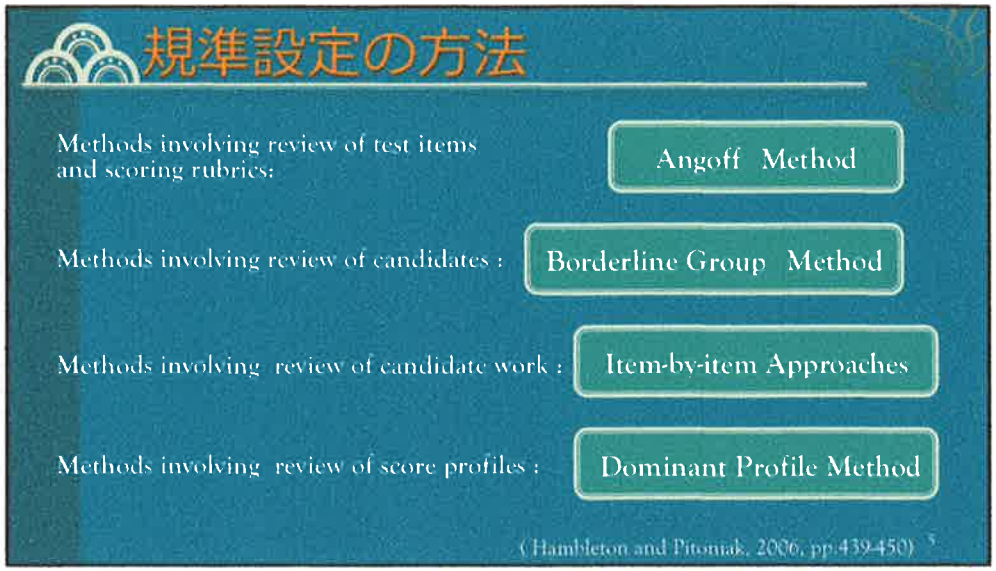
私どもの研究課題は、ごく簡単に言えば、「英語のテストが70点であれば、その受験者を、不合格ではなく、合格としてよいのか」 ということを考えてみようということです。

この英文は、American Educational Research Association, APA, NCMEが2014に出版した、あの有名な Standards for Educational and Psychological Testingの中の一節です。たとえば、テスト項目の70%に正解したならば、それを合否を定める分割点としている組織がしばしばある様です。しかし、もしそれが恣意的に定められた分割点であるとすれば、それは次のような理由から問題である、と述べています。その理由は、要約すると、そのテストについての詳しい情報が示されていなければ、そのような分割点は、まったく意味をなさないということです。

私どもの研究課題は、ごく簡単にいえば、「英語70点であれば、合格としてよいのか？」という様な問に答えようとするものです。そして、それでよければ、なぜ良いのか、その理由を考えてみようというものです。

＜規準と規準設定＞

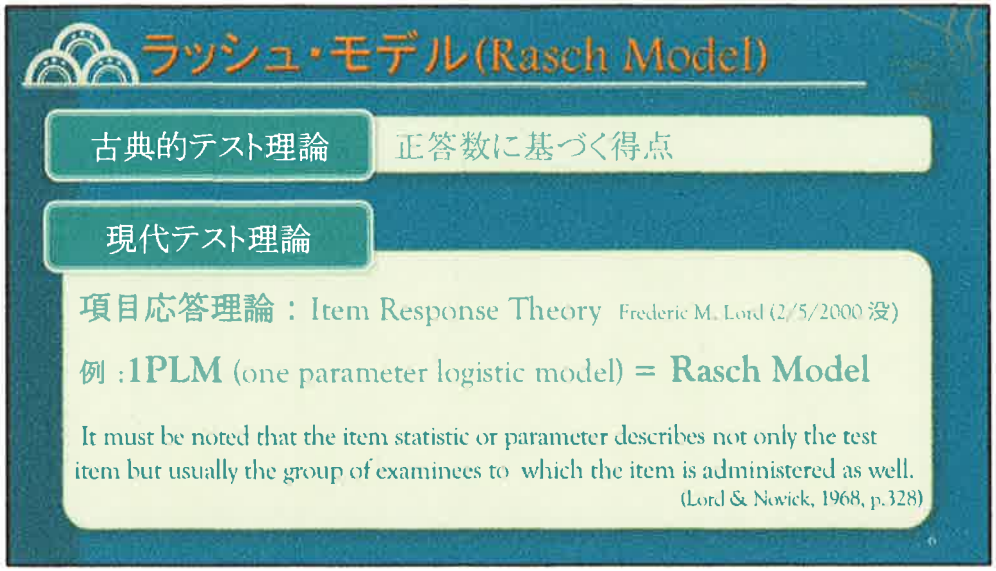
この研究課題の中に示した「規準」と言うのは、英語では、standard, passing score, cutscore, cutoff score, performance level などと言われていますが、これの意味を明確にしておく必要があります。「規準設定」は、英語でstandard setting と言われますが、その意味は、“the process by which a standard or cut score is established” (Cizek ,2006,p.226) で表すことができます。設定した目標に学習者が到達したかどうかをきめる手順ということです。そして、到達と未到達を決定する「分割点」(cut score)をどうしたら、最も適切に設定することができるかを考えることです。



＜規準設定の方法＞:

この「規準設定」の方法に関する研究は沢山行われてきています。例えば、Hambleton & Pitoniak (2006, pp.439-450)では、それを、4つの分野に分けて詳しく説明しています。その具体的名称としては、(1)Angoff Method, (2) Borderline Group Method, (3)Item-by-item approaches, (4) Dominant Profile Method, などがあります。よく耳にするものとしては、Nedelsky method, Bookmark Method などがあります。

こうした方法に対する意見も様々です。規準設定法に対する否定的な見方、中立的な見方、肯定的見方など様々な意見があります。 例えば、no best standard setting としている見方, データ収集には主観的な要素が入るが決定した規準は極めて客観的で重要であるとする見方、 それから、肯定的な見方などがあります。

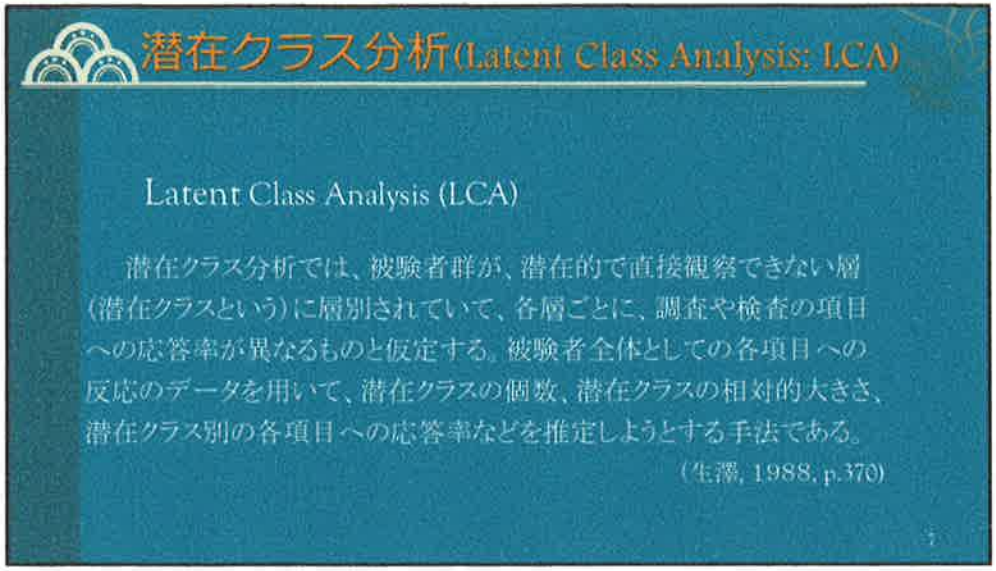


＜Rasch Model：RM＞

英語能力の度合いを、テストでの正答数を基にした点数：素点で表すのが、普通に行われています。山田君の点数は、35点であるとします。しかし、これだけではテストの問題が難しすぎたので35点となったのか、それとも山田君の能力が低いために35点となったのか、という質問に答えることはできますか？できませんね。このように、古典的テスト理論の考え方の一端は、正答数に基づく得点(number-right/correct score) と言われるもので示すことができます。

こうした問題点を鋭く指摘し、新しいテスト理論「項目応答理論」(Item Response Theory: IRT)の開発に着目したのは、Frederic M. Lordと呼ばれる方です。この方は、2000年2月5日に亡くなっています。古典的テスト理論に対して問題点を指摘した発言はたくさんあります。例えば、Lord & Novick (1968, p.328)の画面に出ているような発言があります。

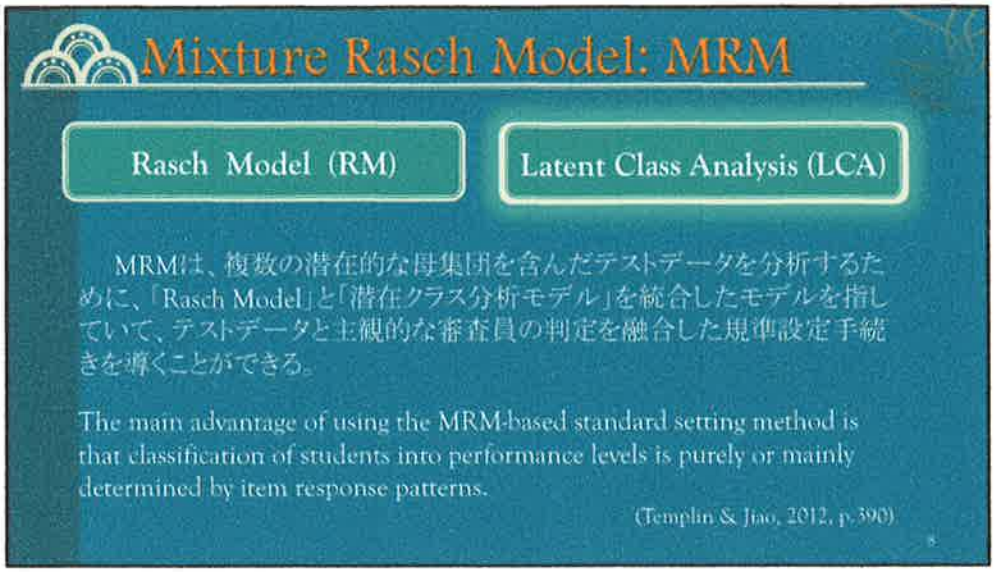
このように、項目応答理論の1つの特徴は、個人の能力特性と使用したテスト問題の項目特性をそれぞれ独立して扱おうとしているものです。そのモデルはいくつも考えられています。そのうちの一つ1PLM(one parameter logistic model) と言うものがあります。これは、デンマークの数学者　Georg Rasch の考案したモデルと同じということがわかり、それをRasch Model と一般に呼ぶことになったわけです。



＜潜在クラス分析（latent Class Analysis: LCA）＞

これから述べようとしている「混合ラッシュモデル」（Mixture Rasch Model: MRM）は、このRasch Model と「潜在クラス分析」（Latent Class Analysis: LCA）とを用いたモデルであると言われています。したがって、ここでは、「潜在クラス分析」とは何かという概念の大枠を把握しておくことが必要です。「潜在クラス分析」は、Lazarsfeld, P.E. (1950). によって紹介されたものです。その詳細に関しては、ここでは省略しますが、ごく簡単にいえば、次の様なことです。

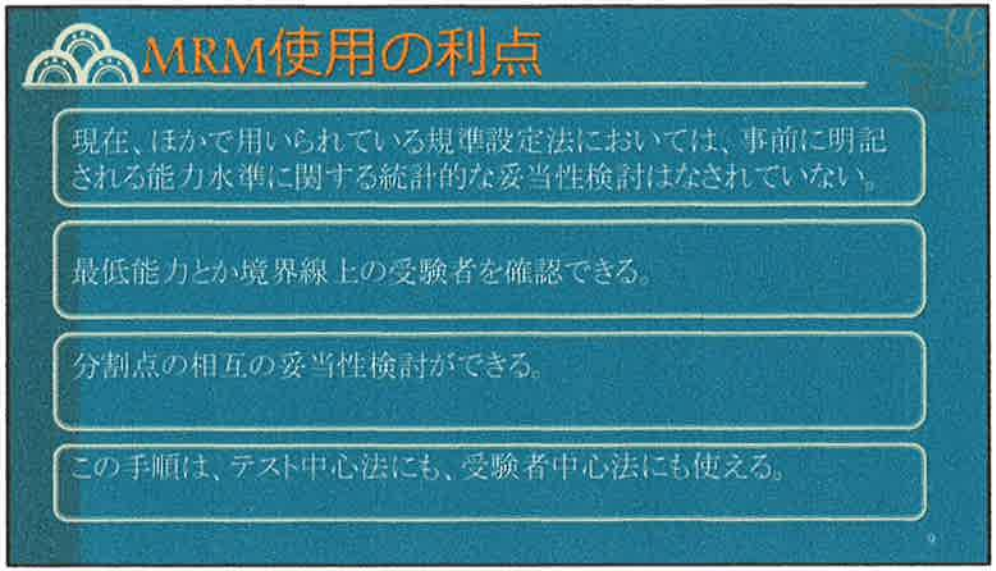
潜在クラス分析では、被験者群が、潜在的で直接観察できない層（潜在クラスという）に層別されていて、各層ごとに、調査や検査の項目への応答率が異なるものと仮定する。被験者全体としての各項目への反応のデータを用いて、潜在クラスの個数、潜在クラスの相対的な大きさ、潜在クラス別の各項目への応答率を推定しようとする手法である。（生澤, 1998, p.370）



＜ Mixture Rasch Model :MRM ＞

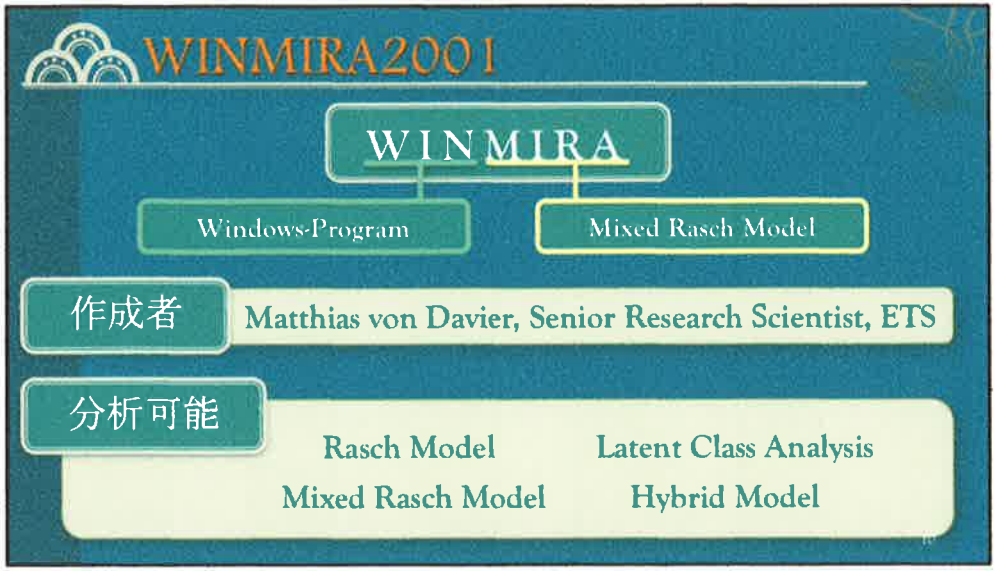
Mixture Rasch Model：MRM (混合ラッシュモデル)というのは、複数の潜在的な母集団を含んだテストデータを分析するため、Rasch Model と潜在クラス分析モデルを統合したモデルを指していて、テストデータと主観的な審査員の判定を融合した規準設定手続きを導くことができると考えられています。

しかし「混合ラッシュモデル」を利用することの重要な、そして有利な点は、能力水準に関する受験者の分類は、主に「項目の反応形式」によって決定されるということです。もし、それが理想通りに行われるのであれば、これまでの古典的規準設定モデルにおける人間による判断と比べても、その分類の誤差は、かなり減少されることになるでしょう。そう言うことが、つぎの(Templin & Jiao, 2012, p.390)で言われています。



＜Mixture Rasch Model: MRMの利点＞

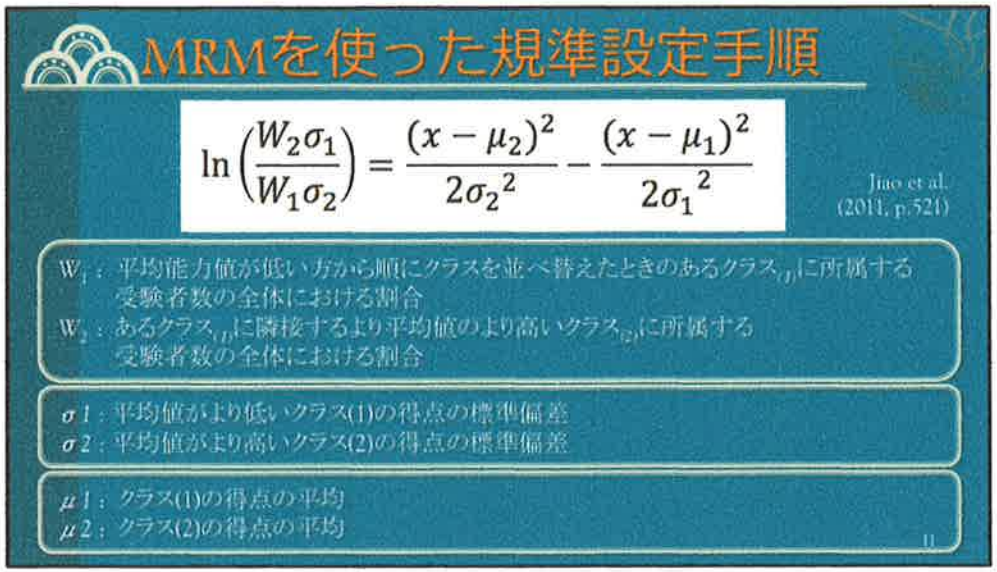
1). これまで用いられている規準設定法においては、事前に明記される能力水準に関する統計的な妥当性検討はなされていない。
2). 最低能力とか境界線上の受験者を確認できる。
3). 分割点の相互の妥当性検討ができる。
4). この手順は、テスト中心法にも、受験者中心法にも使える。 (Jiao et al. 2011, p.514)



＜WINMIRA2001＞とはどんなソフトか：

MRMに使われるソフトの一つに、WINMIRAというものがあります。この名称はWINMIRA といいますが、WIN は WINDOWS-programのWIN, MIRAは、Mixed Rasch Model のMIRAを組み合わせたものです。さきにMixture Rash Model と言いましたが、このように、Mixed Rash Model と呼ぶ場合もあります。

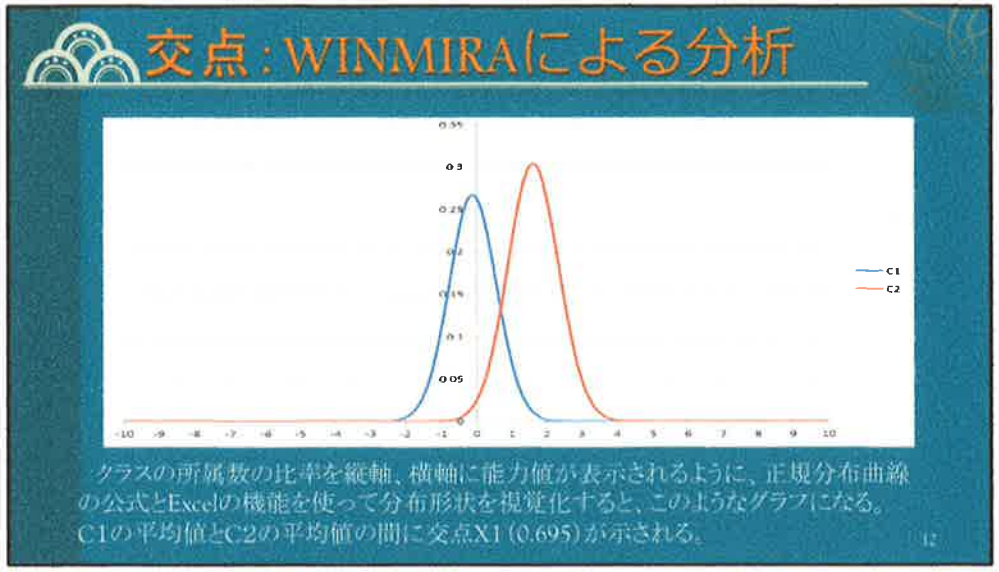
このソフトの作成者は、Matthias von Davier (2001)という方ですが、アメリカのテスト研究開発で有名なETS (Educational Testing Service) の主任研究員（Senior Research Scientist）をした方です。そして、このソフトは、 Rasch Model, Latent Class Analysis, Mixed Rasch Model, Hybrid Model の分析・研究に使われてきています。



＜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ますが、MRMを使った規準設定の研究は、それほど多くはありません。例えば、Jiao et al.(2011)は、10,000人の読解テスト40項目で5クラスに分割するためにMRMを使っています。Kreiner. (2007)もMRMを使用した分析法の一つとして注目を浴びています。また、大友、中村、法月(2016)では、語彙テストを使って、50問、213名のデータ分析にこのMRMを使ってみました。

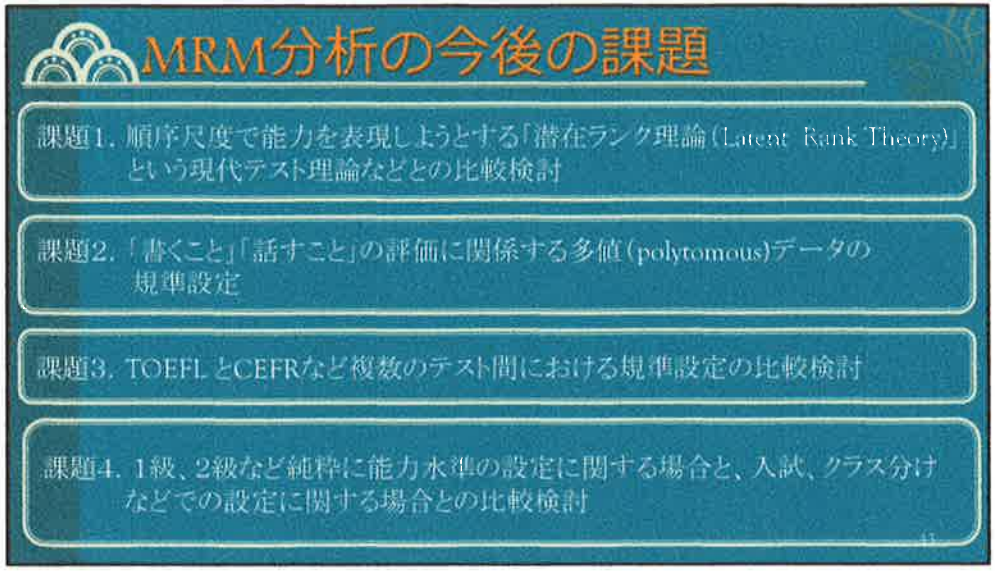
この画面は、大友、中村、法月（2016, pp.37-38）に関する計算手順の一端で、分割点設定手順の準備で使われているものです。



＜交点：WINMIRA による分析＞

＜. WINMIRA による分析＞ ここでは、C1とC2の平均値の値に位置する X1 = 0.695, C1正規曲線と2正規曲線の左側の裾で交わる地点X2 = −24.543を求めることができます。

クラスの所属数の比率を縦軸、横軸に能力値が表示される様に、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、この様なグラフになります。C1の平均値（−0.100）とC2の平均値（1.621）の間に交点X1 (0.695)が示されるようになります。



< MRM分析の今後の課題 >:

残されている今後の課題は多いが、そのうちのいくつかをあげれば、次の4点になります。

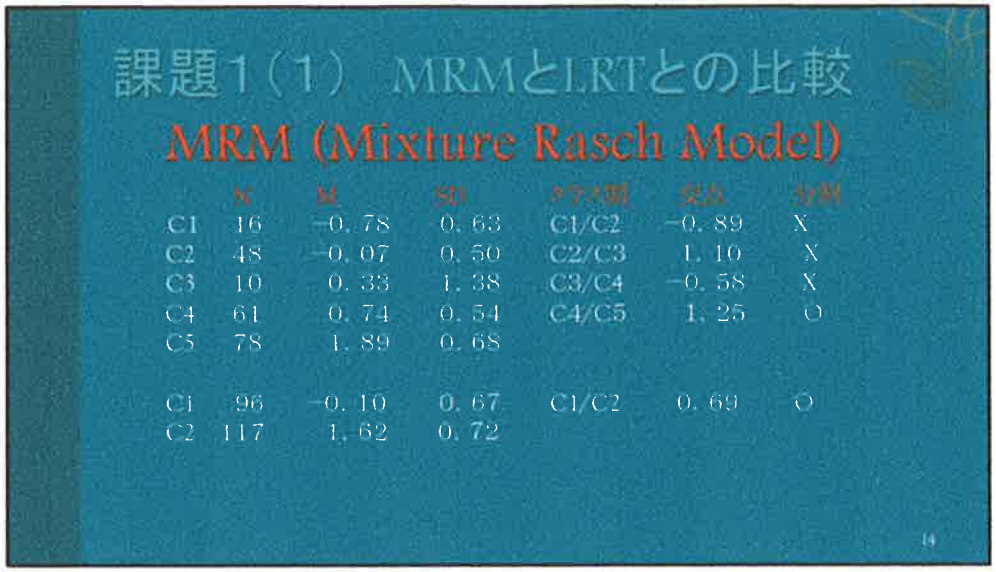
課題1. MRM法と潜在ランク理論(LRT)との比較検討:

課題2. 多値データの規準設定:

課題3. CEFR と TOEFL など複数の規準の比較検討:しかし、しばしば、“cut scores are constructed, not found” (Zieky, 2001, p.45) を再考させる結びになっている場合に巡り会います。

課題4. 分割点の利用に応じた規準設定:

1級、2級といった純粋に能力水準の設定に関する場合と、入試、クラス分けなどに関する設定に関する場合とは、混同せずに、完全に切り離した基準設定の考察が必要なのでしょうか。



＜課題1（1） :MRM（Mixture Rasch Model）＞

先に示した「今後の課題」は、目下検討中ですが、その一端を、紹介いたします。
課題1に関連しているMRM（Mixture Rasch Model）とLRT (Latent Rank Theory) による分析結果を比較検討したものです。

まず、課題1（1）では、
WINMIRA 2001 を使って、日本人大学生213名に実施した50問の英語語彙テスト結果を、分析したものです。5クラス分析と2クラス分析における各クラスのN＝受験者数、M＝平均点、SD＝標準偏差、を示します。さらに、Jiao, H . et al (2011) が提案した数式に基づき、隣接するクラス間で交わる地点となる分割点を計算したものです。例えば、5クラス分析のC1とC2の交点は、-0. 89であったが、この値は、C1の平均値-0. 78よりも低く、両クラスの平均値間に位置していないため、2つのクラスを分割する妥当な地点ではないと言えます。したがって、分割はXとしてあります。同じことが、C2/C3, C3/C4にも見られます。

MRM2クラス分析の交点0. 69は、実測値で比較すると、0. 58（素点29点） と0. 70（素点30点）の間に位置します。

<課題1（2）LRT(Latent Rank Theory)>

同じデータ213名に実施した50問の語彙テストの結果です。そのデータのRasch Model による受験者能力推定値を、WINSTEPS(Linacre,2014)と呼ばれる統計ソフトで算出しました。それを、潜在ランク理論（Latent Rank Theory:LRT) の分析ソフトである Exametrika (荘島(Shojima), 2011)を使って各受験者にランクをつけました。そして、前に述べたMRMと同じ方法で、隣接するランク・グループの交点を計算してみたものです。

この課題1（2）では、このLRT法の5ランク分析、2ランク分析いずれのランク間でも有効な分割点設定を行うことが、可能でした。

LRTによる2クラス分析の交点−0．06は、実測値で比較すると、-0.03（素点31点）と-0.16（素点30点）の間に位置します。

＜今後の課題（1）確認＞

今後の課題は、4つほど掲げていますが、そのうちの一つの中間報告の確認事項です。今後の課題1. は、順序尺度で能力を表現しようとする「潜在ランク理論」という現代テスト理論などとの比較検討についてです。

その確認事項は、「混合ラッシュモデル分析とは異なり、潜在ランク理論法で分析したすべてのランク間においては、有効な分割点が得られることが確認できた」ということです。

その分析結果をさらに比較・検討すると、サンプルの受験者数がそれほど多くはないことも要因として考えられますが、すくなくともこのテストデータでは、Jiao et al (2011) が提唱する分割点設定法は、「 混合ラッシュモデル（MRM）分析」よりも「潜在ランク理論（LRT）法に適しているようです。

残された「今後の課題」は、たくさんあります。現在、中村・法月教授を中心に検討・確認を続けております。今後とも、ご指導・ご協力、よろしくお願いいたします。

<参考文献>

参考文献としては、次のようなものがあります。ご清聴、有難うございました。


参考文献
Linacre, M. (2014). WINSTEPS Rasch measurement computer program (Version 3.81.0). Winsteps.com.
Lord, F.M. & Novick, M.R. (1968). Statistical theories of mental test scores (p.328). Addison-Wesley.
Ohtomo, K. (2015). The Founding Father of Language Testing in Japan. In A. J. Kunnan (Ed.) Talking About Language Assessment (pp.103-125). Taylor & Francis.
Templin J. & Jiao, H. (2012). Applying model-based approaches to identify performance categories. In G.J. Cizek (ed.) (2012). Setting performance standards: second edition. (p.390). Routledge.
Von Davier, M. (2001). WINMIRA (Computer software). ASC Assessment System Cooperation.
Zieky, M.J. (2001). So much has changed: How the setting of cutscres has evolved since the 1980s . In G.J. Cizek (ed.) Setting performance standards: Concepts, Methods, and Perspectives. (p.45). Lawrence Erlbaum Associates, Publishers.
大友賢二・中村洋一・吉月 健(2016)『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』(pp.37-38).公益財団法人 日本英語検定協会.
生澤雅夫(1998) 潜在構造分析. 東・梅本・芝・梶田(編)『現代教育評価事典』(p.370) 金子書房.
荘島宏二郎(2010)「ニューラルテスト理論」植野真臣・荘島宏二郎(編)『学習評価の新潮流』朝倉書店
荘島宏二郎(2011) Exametrika (Verson 5.3) [Software]
18

＜参考文献＞

THANK YOU
FOR YOUR ATTENTION

ohtomokenji@gmail.com

19