英語 4 技能テストにおける受験者能力の推定：TOEIC L&R・S&W および TOEIC Bridge L&R・S&W のデータを用いて

卯城祐司[1]・小室竜也[2]

1. 日本国際学園大学教授・筑波大学名誉教授

2. 東北大学・日本学術振興会 特別研究員 PD

目次

1. はじめに

本研究の目的は、TOEIC ® Listening and Reading (以降 TOEIC L&R)、TOEIC ® Speaking and Writing (以降 TOEIC S&W)、TOEIC Bridge ® Listening and Reading (以降 Bridge L&R)、TOEIC Bridge ® Speaking and Writing (以降 Bridge S&W) における 4 技能間の関係性を検討することである。それぞれのテストの問題数および問題形式は表 1 に示される通りである。問題形式および問題数、受験方法などが異なることから、4 技能の関係性を検討する際には、平均点のみを使った分析では誤った結論に至ってしまう。この問題を解決するために、これまでの報告書 (Liao et al., 2010; Liu & Costanzo, 2013) では等化されたスコアを用いられてきた。本研究は先行研究と異なり、等化前のスコアや評価 (本稿では「正解・不正解」の情報およびパフォーマンス評価における「採点・評価情報」のことを指す) に基づき、4 技能の関係を検討する。具体的には、テスト間で異なる問題が用いられている点や受験する技能の違いをより詳しく検討することを目的として、受験者能力を同一尺度上で比較することができる一般化線形混合効果モデル (Generalized Linear Mixed Effect model; GLMM)[1] を用いた。

なお、今回の調査[2]では日本の英語学習者のみを対象としており、等化前のデータや異なる分析手法を用いている。そのため、これまでの先行研究と得られる結果が異なる点には注意が必要である。

表 1

TOEIC L&R・TOEIC S&W・TOEIC Bridge L&R・TOEIC Bridge S&W における問題形式のまとめ

テスト	問題番号	問題数	問題形式	備考
TOEIC Listening	Part 1	6	写真描写問題	
	Part 2	25	応答問題	
	Part 3	39	会話問題	
	Part 4	30	説明文問題	
TOEIC Reading	Part 5	30	短文穴埋め問題	
	Part 6	16	長文穴埋め問題	
	Part 7	29	1 つの文書	
	Part 7	25	複数の文書	
TOEIC Speaking	QUESTION 1〜2	2	音読問題	採点スケール：0〜3
	QUESTION 3〜4	2	写真描写問題	採点スケール：0〜3
	QUESTION 5〜7	3	応答問題	採点スケール：0〜3
	QUESTION 8〜10	3	提示された情報に基づく応答問題	採点スケール：0〜3
	QUESTION 11	1	意見を述べる問題	採点スケール：0〜5
TOEIC Writing	QUESTION 1〜5	8	写真描写問題	採点スケール：0〜3
	QUESTION 6〜7	2	E メール作成問題	採点スケール：0〜4
	QUESTION 8	1	意見を記述する問題	採点スケール：0〜5

[1] マルチレベルモデル (multi-level model) や階層モデル (hierarchical model) とも呼ばれる統計モデルである。詳細は Brysbaert and Debeer (2025) や Gries (2021)、Linck and Cunnings (2015) などを参照のこと。

[2] Source: Derived from data provided by ETS Copyright © 2024 ETS. www.ets.org。The opinions set forth in this publication are those of the author(s) and not ETS. 本研究において ETS はデータ提供のみを行っており、研究の設計、分析、解釈、ならびにその結果に関する一切の責任は著者らにあります。

TOEIC Bridge Listening	Part 1 (Q1〜6)	6	画像選択問題	
	Part 2 (Q7〜26)	20	応答問題	
	Part 3 (Q27〜36)	10	会話問題	
	Part 4 (Q37〜50)	14	説明文問題	
TOEIC Bridge Reading	Part 1 (Q51〜65)	15	短文穴埋め問題	
	Part 2 (Q66〜80)	15	長文穴埋め問題	
	Part 3 (Q81〜100)	20	読解問題	
TOEIC Bridge Speaking	QUESTION 1〜2	2	音読問題	採点スケール：0〜3
	QUESTION 3〜4	2	写真描写問題	採点スケール：0〜3
	QUESTION 5	1	聞いたことを伝える問題	採点スケール：0〜3
	QUESTION 6	1	短い応答問題	採点スケール：0〜3
	QUESTION 7	1	ストーリー作成問題	採点スケール：0〜4
	QUESTION 8	1	アドバイスをする問題	採点スケール：0〜4
TOEIC Bridge Writing	QUESTION 1〜3	3	文を組み立てる問題	採点スケール：0〜2
	QUESTION 4〜6	3	写真描写問題	採点スケール：0〜3
	QUESTION 7	1	短文メッセージ返信問題	採点スケール：0〜3
	QUESTION 8	1	ストーリー記述問題	採点スケール：0〜3
	QUESTION 9	1	長文メッセージ返信問題	採点スケール：0〜4

2. TOEIC L&R および TOEIC S&W の分析

分析の対象となるデータは 3957 名分であり、採点期間は以下の通りである。

- TOEIC S&W：2023 年 4 月 1 日〜2024 年 3 月 31 日
- TOEIC L&R：2022 年 10 月 1 日〜2024 年 9 月 30 日

なお、今回の分析データは団体特別受験制度[3]を利用した受験者を対象としている。また、一定期間内に複数回受験の場合は、最も受験期間が短いデータを採用した。S&W は技能別受験を可能としているため、以下の 3 つの欠損パターンが生じている：(a) スピーキングとライティングの両方のスコアがある、(b) スピーキングのスコアのみがある、(c) ライティングのスコアのみがある。これらの欠損値の扱いについては、R (ver. 4.5.1; R Core Team, 2025) のパッケージである lme4 (ver. 1.1-36; Bates et al., 2015) および ordinal (ver. 2023.12-4.1; Christensen, 2019) によるリストワイズ削除 (listwise deletion) を行った。そのため、最終的な分析対象は 3201 名となる。

本研究で分析の対象としたテストフォームは TOEIC L&R で 24 種類、TOEIC S&W で 52 種類である。本研究で扱うデータは等化前のスコアを対象としているため、受験者および項目レベルでの詳細な分析を可能とする GLMM を用いた。また、テストフォームによって項目が異なることから、本研究はこれらを別の項目として扱った。例えば、A のフォームにおけるリスニングの第 1 問は、別のテストフォームである B におけるリスニングの第 1 問とは異なる項目として扱われている。

TOEIC L&R はリスニング 100 問、リーディング 100 問で構成されており、それぞれの問題に対して正解か不正解かの 2 値データが得られている。一方、TOEIC S&W はスピーキング 11 問、ライティング 8 問で構成されており、0 から 5 の多値データが得られている。2 値と多値という異なるデータの特徴を有することから、テストの合計点を用いた古典的テスト理論による分析では、これら 4 技能の関係を適切に捉えることが難しい。そこで、

[3] TOEIC Program を団体で申し込みの上受験する制度

本研究では受験者能力という同一尺度の潜在変数を仮定する項目応答理論 (Item Response Theory; IRT) とモデリングが等しい GLMM を用いて、4 技能における関係性を検討した。具体的には、周辺最尤推定および EM アルゴリズムに基づく受験者能力を推定し、共通受験者による推定値を用いて 4 技能間の関係性を検討した。分析に際し、2 値データは lme4 パッケージによる 2 値ロジスティックモデル、多値データは ordinal パッケージによる累積リンク混合モデルをそれぞれ使用した。

2.1 分析の理論的背景

古典的テスト理論 (Classical Test Theory; CTT) では平均値やクロンバックの α などによって、テストの傾向を把握するために古くから用いられてきた。しかし、CTT は受験者および項目の性質に依存している。つまり CTT は一般に、異なる受験者の比較や、異なる時期に実施したテストの比較には向いていない。そこで本分析では IRT を用いる。

IRT はデータにモデルを適用する統計モデリングのアプローチである。具体的には、項目の困難度と受験者の能力という潜在変数から正解・不正解という 2 値データや段階評価である多値データを予測する。項目 j への正答確率p_j (θ) は潜在特性 θ と項目 j の特徴を表すb_j (項目困難度) を用いて以下のように θ の関数で表す。$b_j = \theta$ において正答確率は 0.5 となるため、項目困難度のパラメータは「項目を 50%の確率で正解できる特性値でその項目を特徴づけたパラメータである」と解釈できる。b_jは値が大きいほど難しい項目であり、項目応答関数は右側に位置する。

$$P(\theta) = \frac{\exp(\theta - \mathrm{b}_j)}{1 + \exp(\theta - \mathrm{b}_j)}$$

受験者能力θ_iと項目困難度b_jが所与の全項目への反応ベクトルu_{ij}が得られる確率は以下のように表される。

$$f(u_{ij}|\theta_i,\, b_j) = \frac{exp[u_{ij}(\theta_i - b_j)]}{1 + exp(\theta_i - b_j)} = \frac{exp(r_i\theta_i - \sum_{j=1}^{n} u_{ij}b_i)}{\prod_{j=1}^{n}[1 + exp(\theta_i - b_j)]}$$

ここでr_iは受験者 i の総合点を表し、$r_i = \sum_{j=1}^{n} u_{ij}$と表される。この関数の対数を取り、項目困難度 b と受験者能力 θ を最大化する推定方法は Joint Maximum Likelihood Estimation (JMLE) と呼ばれる。JMLE は受験者能力と項目困難度を同時に推定しているが、現在広く用いられている Marginal Maximum Likelihood Estimation (MMLE) ではこれらの潜在変数を別々に推定している。MMLE は GLMM の推定にも用いられており、本研究では MMLE に基づく EM アルゴリズムを採用している。

目的変数が順序データとなる場合 (例：5 段階評価)、Rating Scale Model (RSM; Andrich, 1978) と Partial Credit Model (PCM; Masters, 1982) が用いられる。正解・不正解の 2 値データにおいて、受験者 i が項目 j に解答する際、正解 (1) の確率$p_{i1}(\theta_i)$と不正解 (0) の確率$p_{i0}(\theta_i)$は以下のように表される (図 1 も参照)。

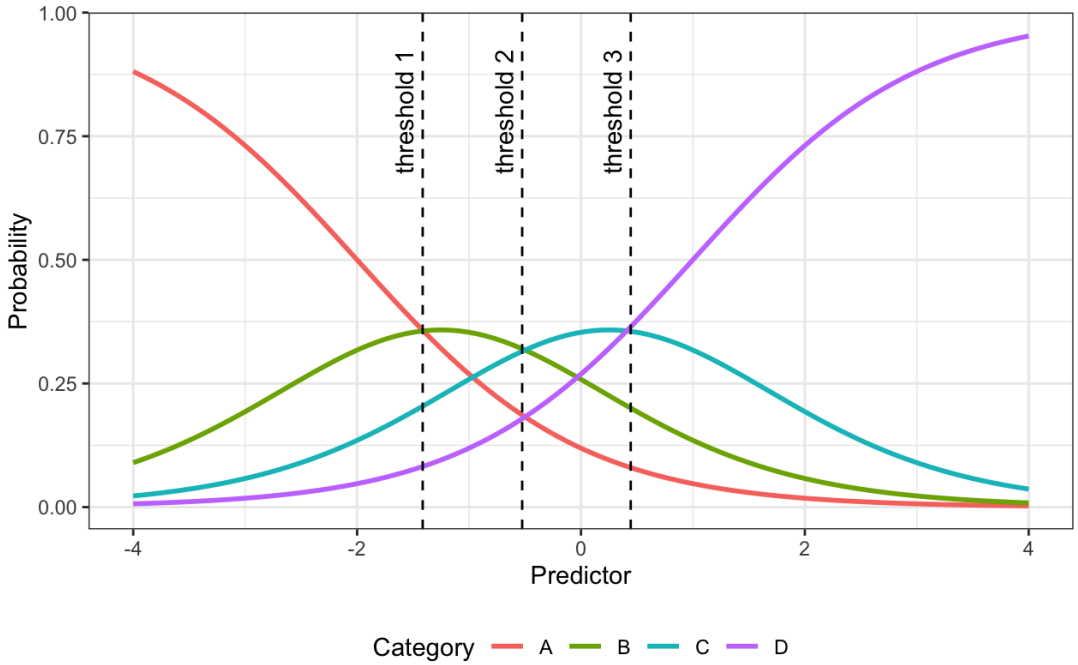
$$\pi_{j10}(\theta_i) = \frac{p_{j1}(\theta_i)}{p_{j1}(\theta_i) + p_{j0}(\theta_i)} = \frac{exp(\theta_i - b_j)}{1 + exp(\theta_i - b_j)}$$

ここで$\pi_{j10}(\theta_i)$は正解 (1) と不正解 (0) という 2 つのカテゴリに反応する確率を指している。RSM および PCM では、このカテゴリを 3 つ以上に拡張している。3 段階評価の場合、RSM では (a) 1 の評価と 2 および 3 の評価の比較、(b) 1 および 2 の評価と 3 の評価の比較という 2 つの比較が行われる。一方、PCM では (a) 1 の評価と 2 の評価の比較、(b) 2 の評価と 3 の評価の比較という 2 つの比較が行われる。5 段階評価の場合には 4 つの比較、

4 段階評価の場合には 3 つの比較のように、C 個のカテゴリの場合には C-1 の比較が行われる。具体的には、以下の図のように、4 つのカテゴリの場合には 3 つ (4 - 1) の比較が行われている。それぞれのカテゴリ曲線が交差する点は遷移点・閾値 (threshold) と呼ばれる。これらの遷移点の定義で、大きく (a) 累積モデル (cumulative model)、(b) 隣接カテゴリモデル (adjacent-category model) に分けられる。本研究では累積モデルによる RSM を採用している。

図 1

4 段階評価におけるカテゴリ特徴曲線



これらの IRT に基づくモデリングアプローチは、マルチレベルのデータを扱う回帰分析の枠組みと共通している。マルチレベルのデータは「生徒-学校」「社員-会社」「町-市-都道府県」のように、階層のある包含関係を持つデータセットのことを表す。テストの文脈では、1 人の受験者が複数の問題を解くことから「項目-受験者」のようなマルチレベルのデータセットであり、GLMM で分析が可能である。実際に、GLMM で仮定される項目と受験者のランダム切片は、それぞれ IRT における項目難易度と受験者能力に該当する (Bürkner, 2020; Dunn, 2024)。なお GLMM による分析においては、推定される値は係数が逆になるため、値が高いほど易しい項目であり、能力の低い受験者であると解釈されることに注意が必要である。

これらの背景を踏まえ、本研究では受験者能力 (ロジットスケール) を共通した尺度として用いて、4 技能テストの関係を検討する。具体的な分析手順および R コードは Supplemental Material (https://osf.io/9xayb/?view_only=b5d1bddb69254304b26c515dd6429b6b) を参照。

2.2 記述統計

記述統計は表 1 および図 2 の通りである。なお、Listening と Reading は項目に正解したかどうかの 2 値データであるため、平均正解率を示している。一方 Speaking と Writing はパフォーマンス評価であることから、各タスクの評価 (採点スケール) の平均値を報告している。なお、パフォーマンス評価は採点スケールがタスク間で異なるなどのデータの特徴が異なるため、記述統計のみに基づいた 4 技能間での比較は難しい。

表 2

TOEIC L&R の記述統計

	Listening			Reading	
Listening Part	*M*	*SD*	Reading Part	*M*	*SD*
1	81.48	18.18	5	64.69	18.31
2	70.97	15.02	6	63.45	20.56
3	69.94	17.38	7：1 つの文書	61.01	20.23
4	65.84	18.67	7：複数の文書	45.69	22.36

注. 正解か不正解かの 2 値データに基づく正解率を表している。Listening の Part 1 は写真描写問題、Part 2 は応答問題、Part 3 は会話問題、Part 4 は説明文問題。Reading の Part 5 は短文穴埋め問題、Part 6 は長文穴埋め問題、Part 7 は 1 つの文書問題と複数の文書問題。

表 3

TOEIC S&W における Writing の記述統計

QUESTION	*M*	*SD*
1	2.59	0.60
2	2.46	0.68
3	2.41	0.76
4	2.10	0.82

注. QUESTION 1～5 は写真描写問題、QUESTION 6～7 は E メール作成問題、QUESTION 8 は意見を記述する問題。採点スケールはそれぞれ QUESTION 1～5 は 0～3、QUESTION 6～7 は 0～4、QUESTION 8 は 0～5。

表 4

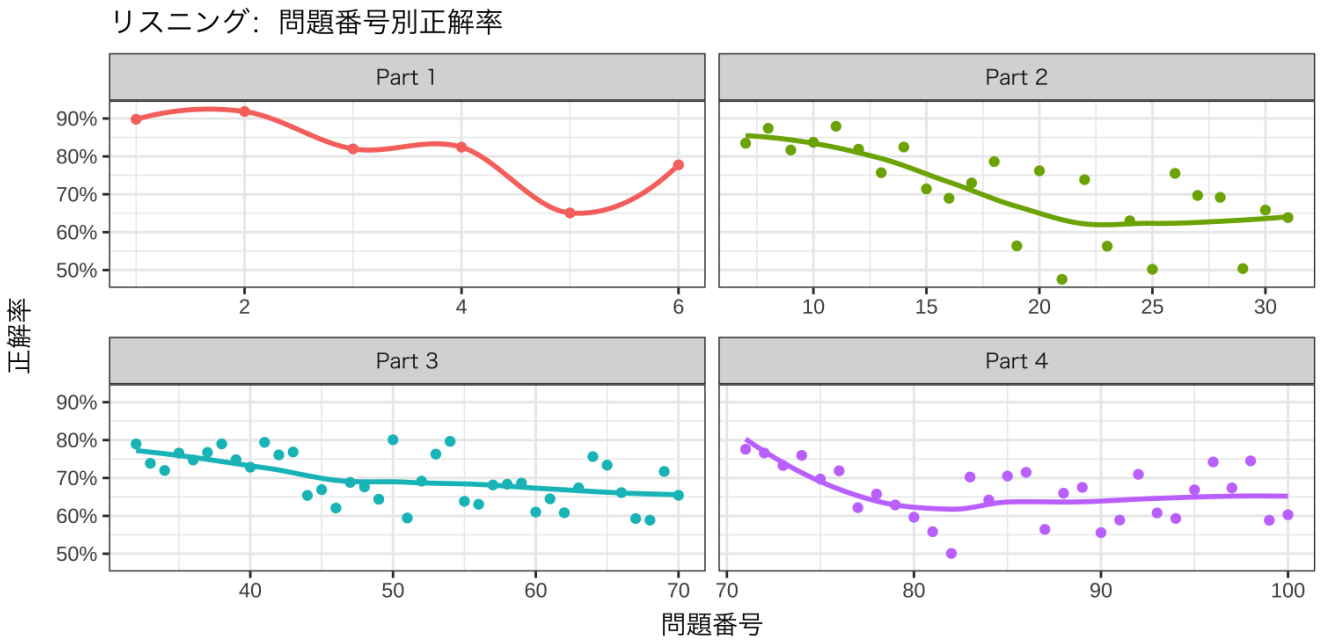
TOEIC S&W における Speaking の記述統計

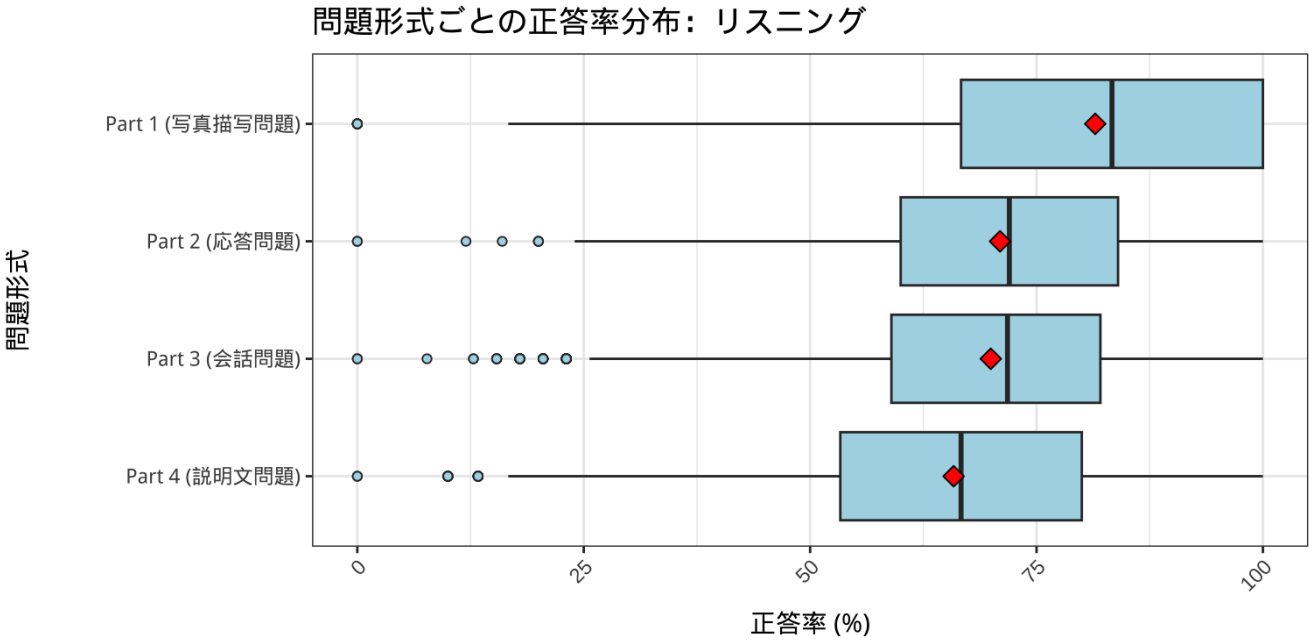
QUESTION	*M*	*SD*
1 (Intonation)	2.11	0.57
1 (Pronunciation)	2.11	0.54
2 (Intonation)	2.16	0.52
2 (Pronunciation)	2.17	0.49
3	2.21	0.56
4	2.17	0.57
5	2.21	0.76
6	2.06	0.78
7	2.16	0.66
8	2.07	0.84
9	1.97	0.87
10	2.14	0.64
11	2.40	0.77

注. QUESTION 1～2 は音読問題、QUESTION 3～4 は写真描写問題、QUESTION 5～7 は応答問題、QUESTION 8～10 は提示された情報に基づく応答問題、QUESTION 11 は意見を述べる問題。採点スケールはそれぞれ QUESTION 1～10 は 0～3、QUESTION 11 は 0～5。

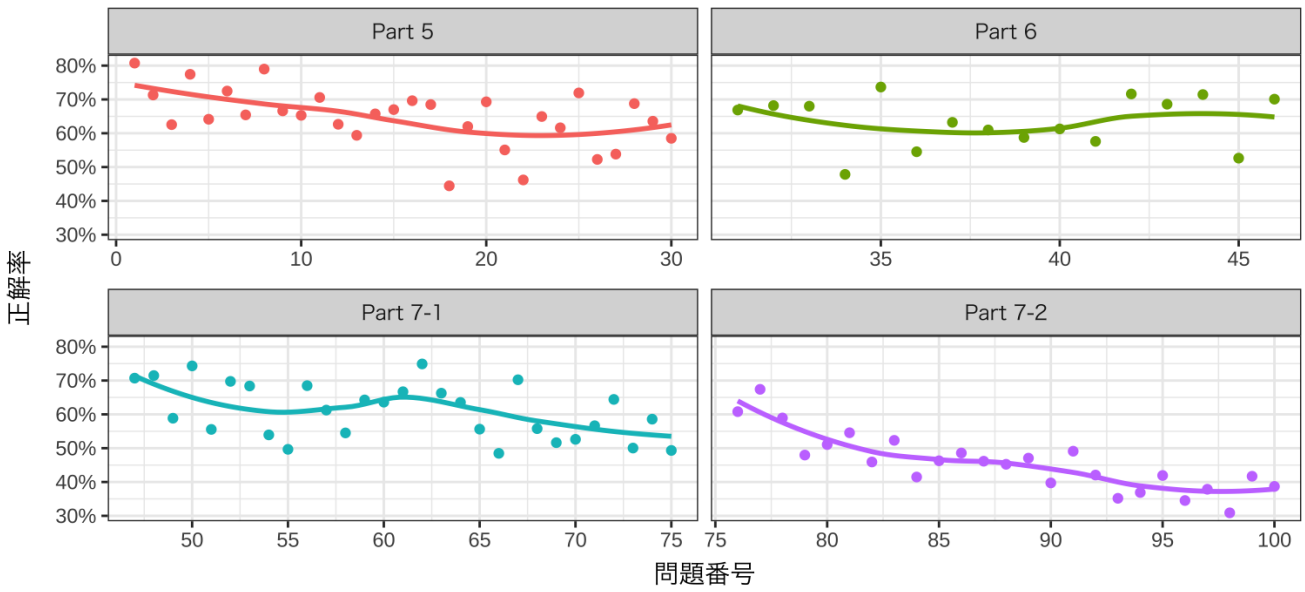
図 2

TOEIC L&R および S&W の記述統計

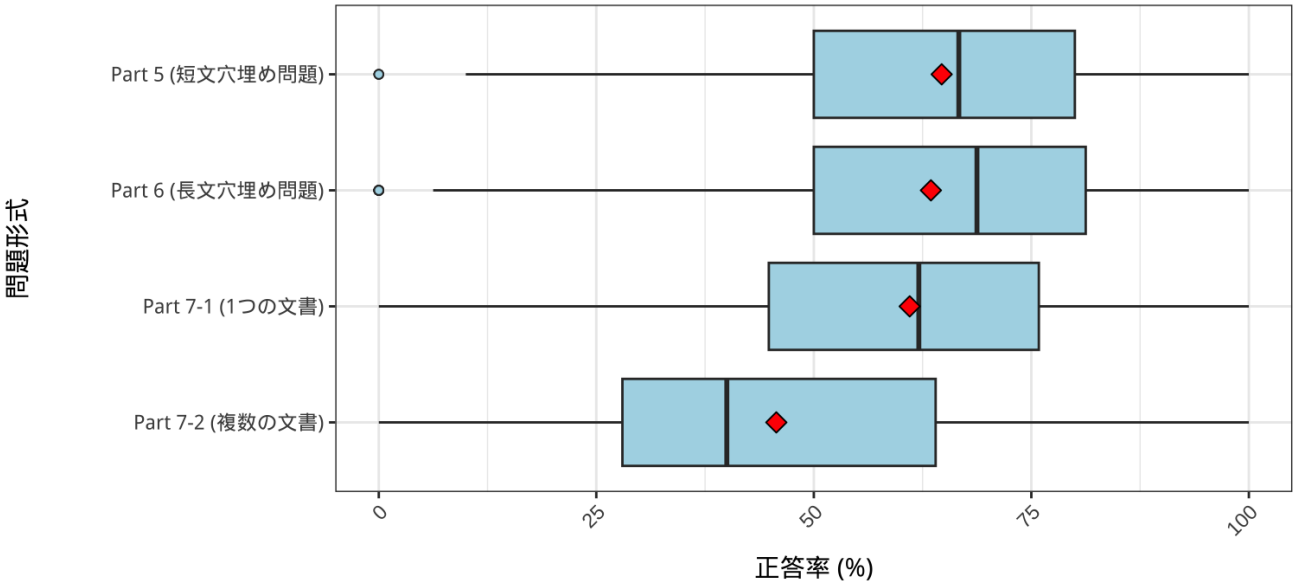


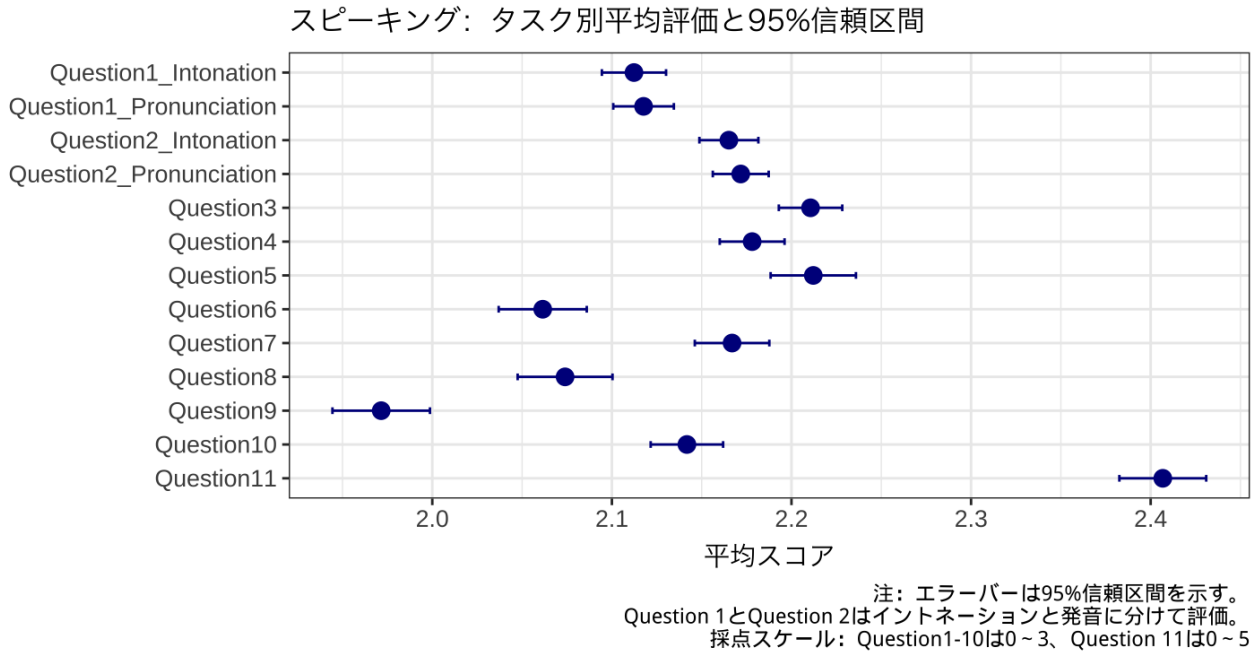


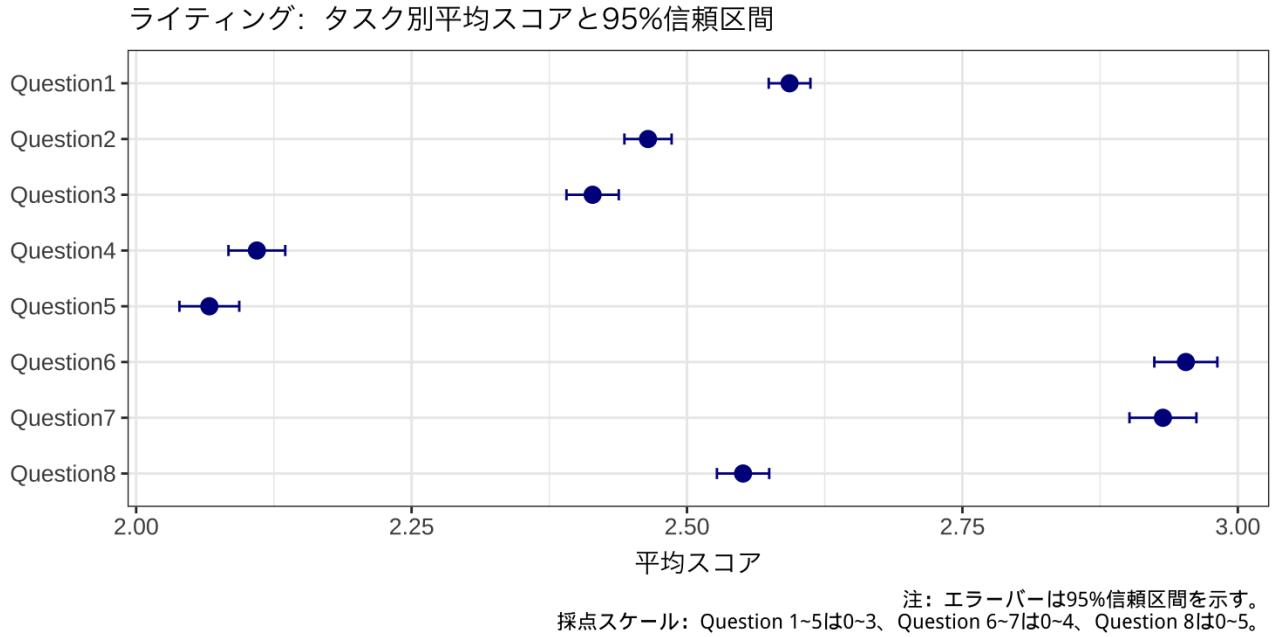
リーディング：問題番号別正解率



問題形式ごとの正答率分布：リーディング







2.3 受験者能力の推定

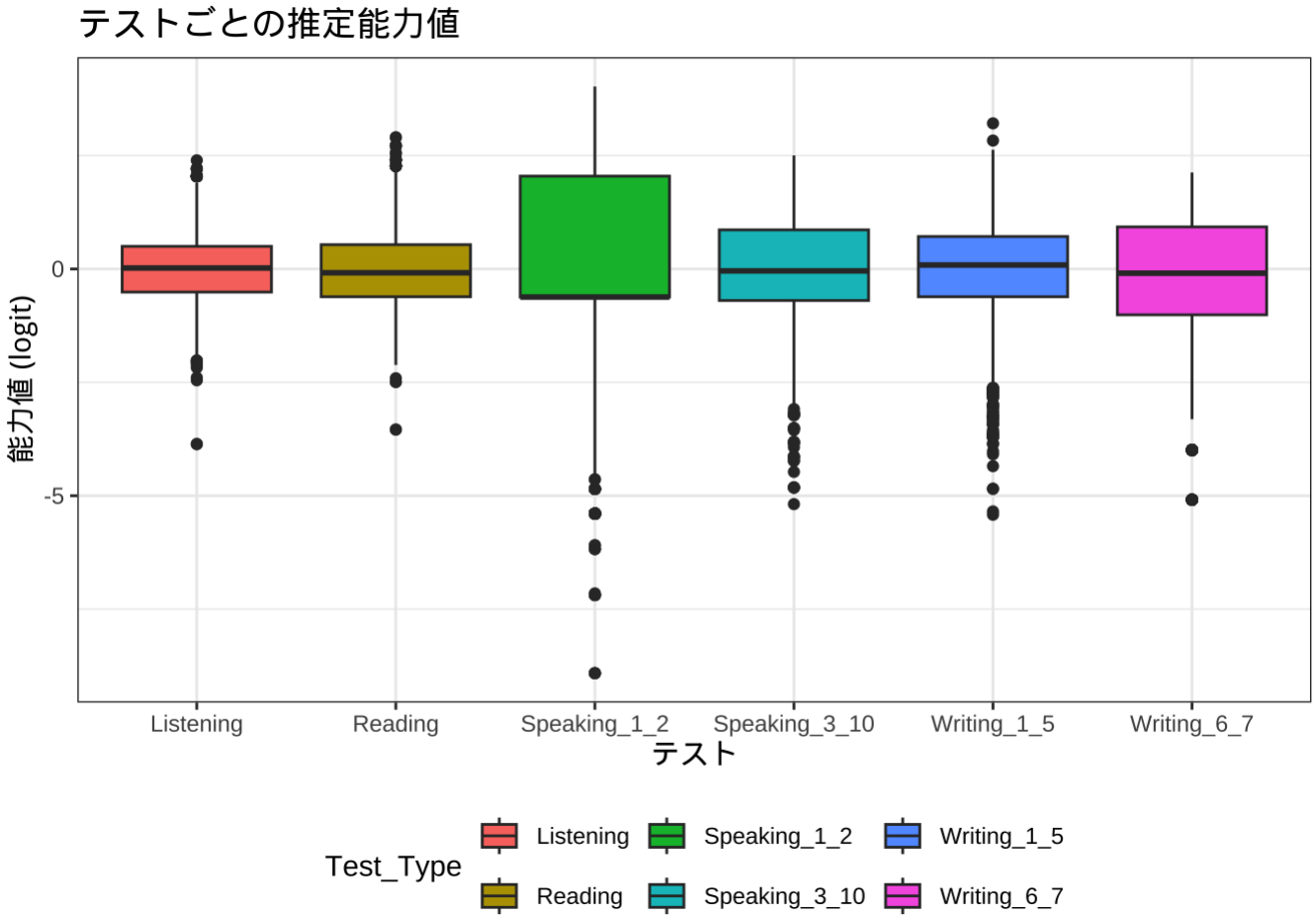
GLMM で推定された受験者能力の結果は図 3 の通りである。Y 軸には受験者の能力がプロットされている。GLMM を採用した最大のメリットは、技能間で異なる採点スケールおよびタスクであったとしても受験者能力という同一の尺度で比較可能とされることである。この受験者能力という潜在変数の値に技能間で有意な差がある場合、例えば、「リーディングよりもリスニングはできない」「スピーキングはできるがライティングはできない」と主張することができる。なお、他のタスクとは採点スケールが異なり 1 水準のみであるスピーキングタスク 11 とライティングタスク 8 は GLMM にて推定できないため[4]、結果は表示されていない。

[4] Question 11 は 0–5 であり、採点スケールが 0–5 の Question は 1 つのみであることを意味している。GLMM では 2 水準以上となるもの (例えば、0–3 の採点スケールは Question1～10 の 10 水準) を対象とするため、今回はスピーキングの Question 11 とライティングの Question 8 は除外されている。

推定値された能力値を目的変数、テスト・タスクを説明変数、受験者を変量効果とした線形混合効果モデルの結果、いずれの技能の間にも有意な差は見られなかった。つまり、特定の技能が突出してできる・できないという傾向は見られなかった。なお、今回得られたテストおよびタスクの間で有意な差がないという結果は、技能間で推定された能力値が等しいことは主張できないことに注意が必要である。

図 3

受験者能力の箱ひげ図



2.4 技能間の受験者能力の相関

本研究の目的は 4 技能の間の関係性を検討することである。これまでの報告書と同様に、相関係数を算出したところ、中から大の正の相関が見られた。特に、リーディングとリスニングは $r = .81$ と高い相関を示した。つまり、リーディングにおける能力が高いほど、リスニングにおける能力も高くなることを意味している (低い場合も同様)。また、スピーキングおよびライティングにおける異なる採点スケールの観点からは、中程度の正の相関が示されている (図 4)。具体的には、スピーキングの採点スケールとして QUESTION 1～10 は 0～3、QUESTION 11 は 0～5 である。0～3 のタスクを潜在変数の受験者能力として扱い Question 11 の結果を予測する場合、分散説明率は 33.6%であった。同様に、ライティングにおける Question 8 の結果の分散は他のライティングタスクで推定された潜在変数からは、33.5%から 39.3%が説明されることが示された。これらの結果から、意見を述べるスピーキングおよびライティング問題は、他のタスクでは測定されない能力が約 6～7 割であった。

リスニングとスピーキングタスクおよびライティングタスクは、中程度の正の相関を示していた ($rs = .49–.68$)。この結果はリーディングも同様であった ($rs = .50–.63$)。概して、英語熟達度が増加するにつれ、4 技能がバラン

スよく成長していく傾向があることが示された。ただし、図 5 に示されるように、必ずしもリスニングで最高点を取った学習者がリーディングやスピーキングで最高の評価スコアを取っているわけではない可能性がある。最低点についても同様に、スピーキングで最も低い評価スコアの学習者であってもリスニングやライティングでは最も低い評価ではない傾向も見られる。つまり、技能間でアンバランスな受験者が見られる可能性がある。これらの傾向から受験者に焦点を当てたクラスター分析を実施した。

図 4

4 技能間の受験者能力の相関係数

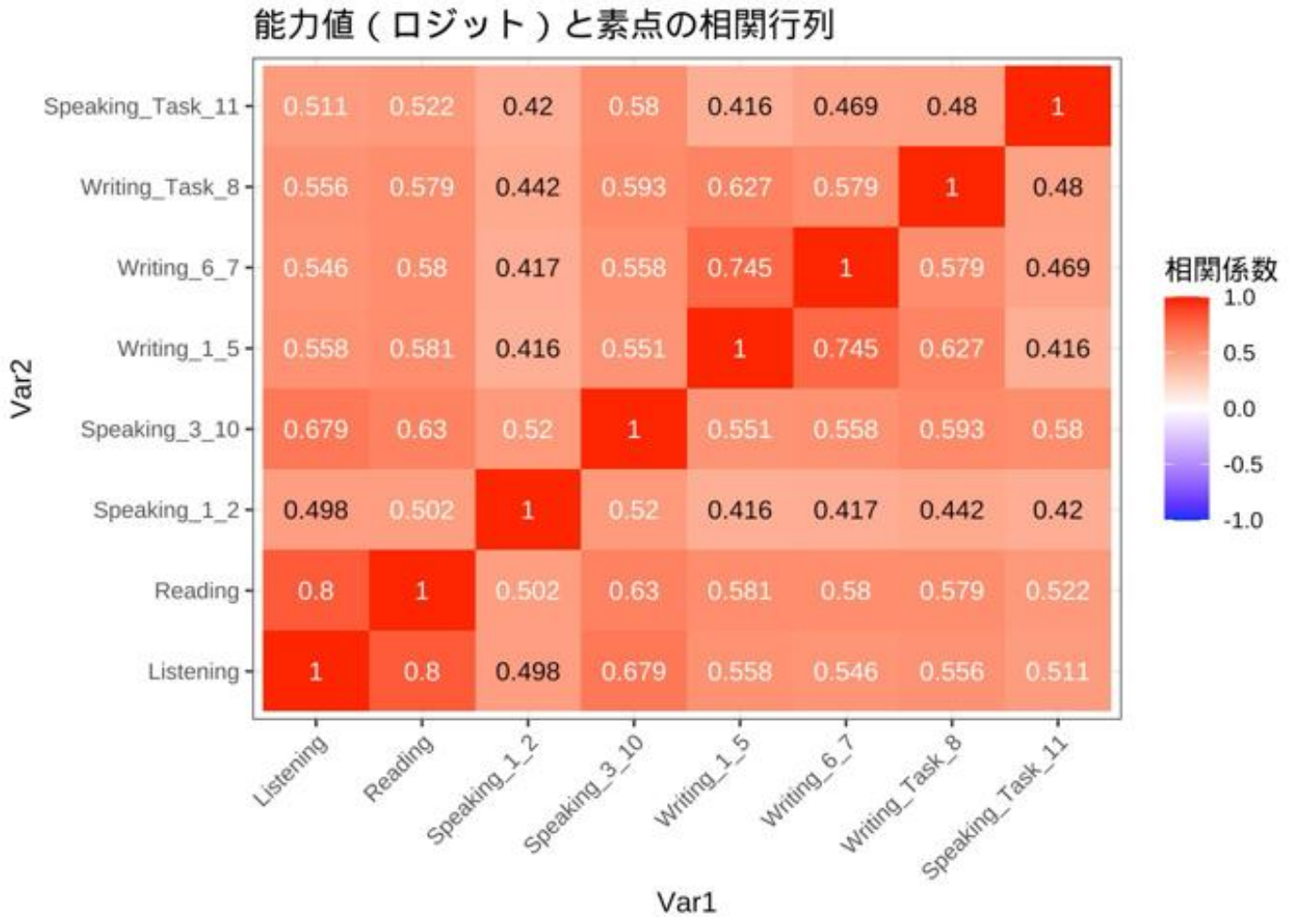
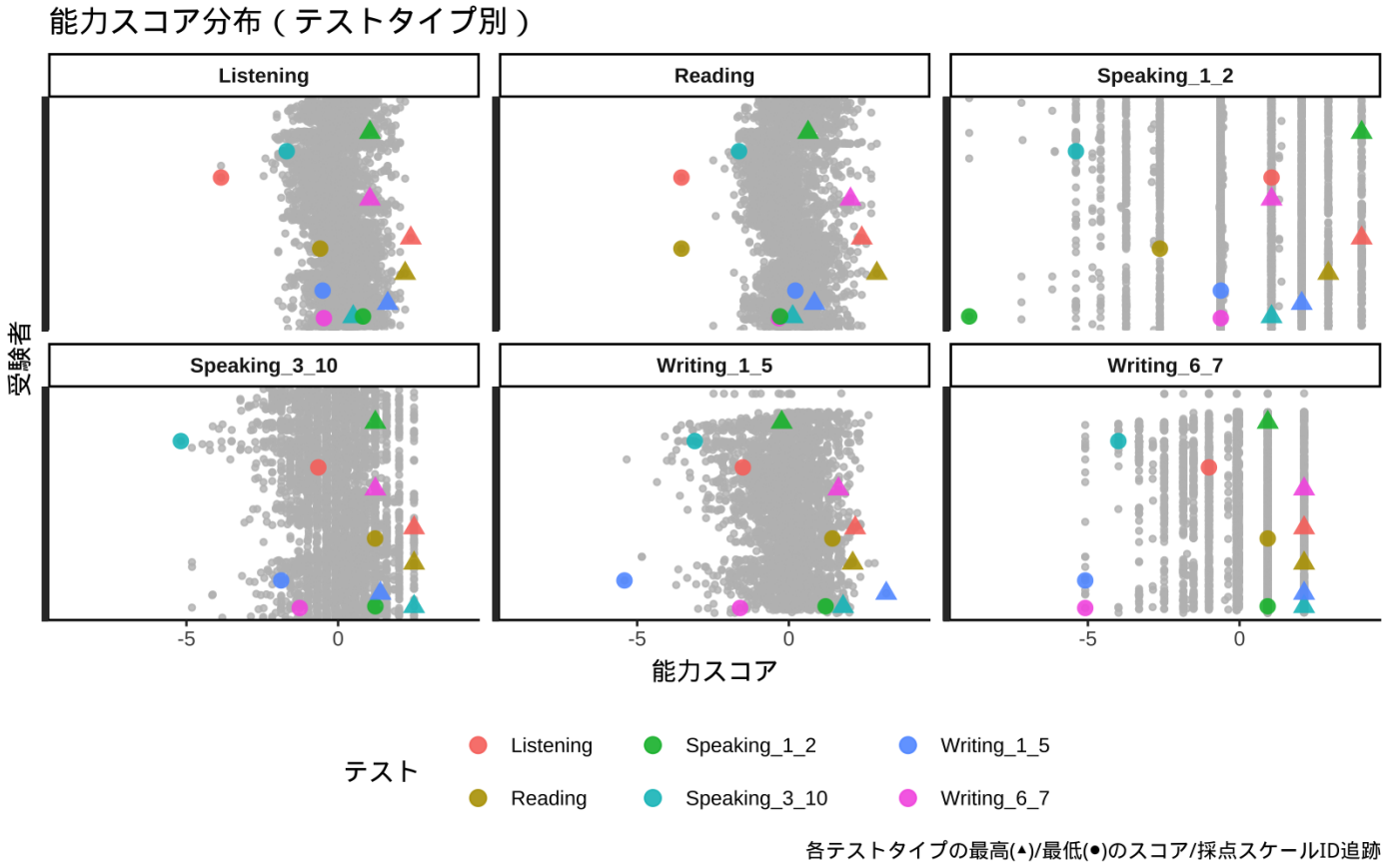


図 5

各テスト・タスクの最低・最高のスコアおよび採点スケールの分布



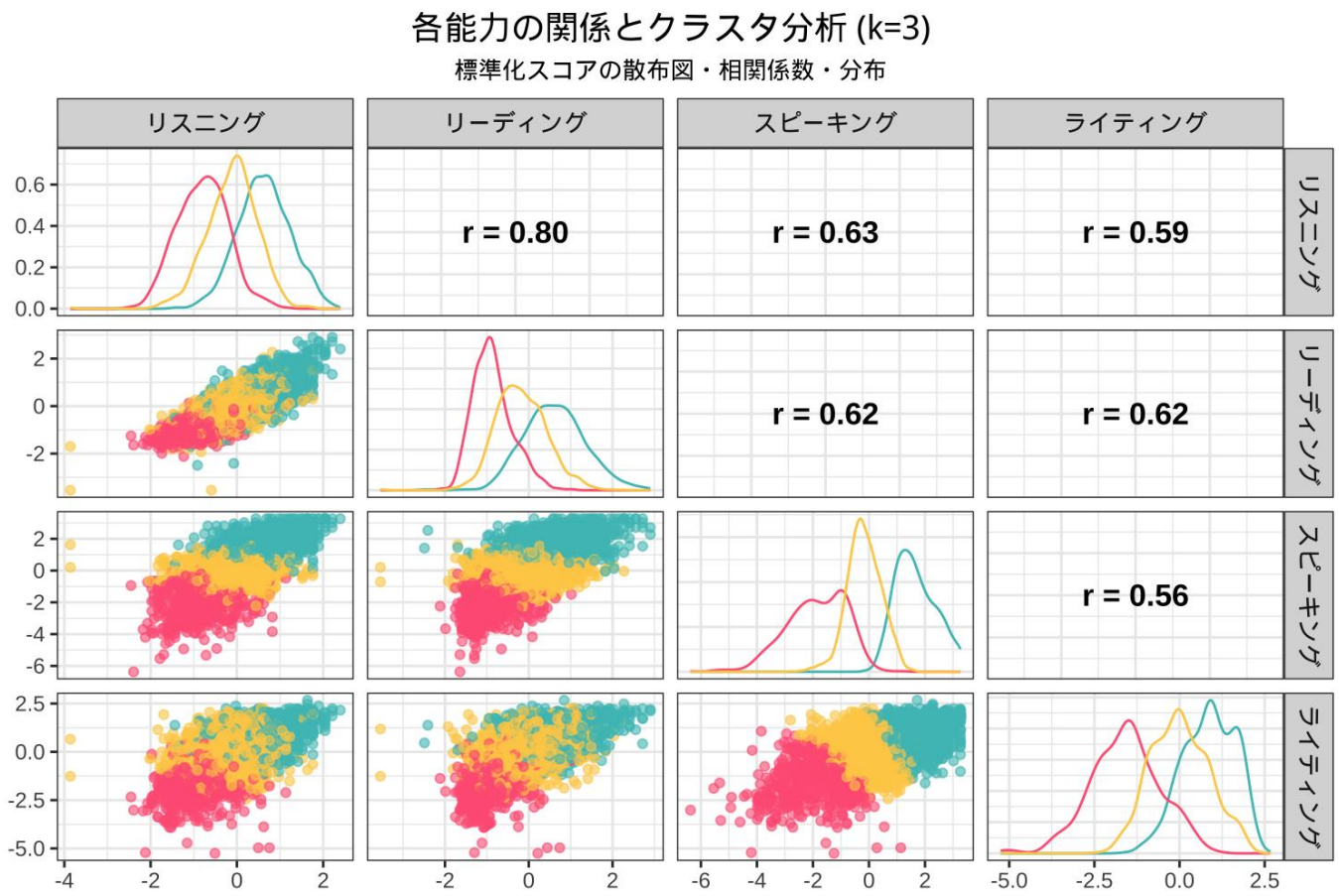
2.5 クラスター分析による能力分布

タスクの間で統計的な差は見られなかったが、受験者によってはリスニングとリーディングは得意で、スピーキングは苦手である英語学習の傾向が見られる可能性がある。そこで、受験者能力をエルボー法および階層的クラスター分析で分類し、傾向をさらに詳しく検討をした（クラスター数決定の手順は OSF を参照のこと：https://osf.io/9xayb/?view_only=b5d1bddb69254304b26c515dd6429b6b)。その結果、受験者は 3 つのクラスターに分けられる可能性が示された（図 6)。タスクにおける推定値を平均化したスピーキングとライティングおよび、リスニングとリーディングの技能ごとの分布は図 6 に示す通りである。この傾向から、低い受験者能力者層（赤)、平均的な受験者能力者層（黄)、高い受験者能力者層（青）の 3 つが検出されたと言える。

4 技能におけるクラスターの能力値の違いは図 7 の通りである。低い受験者能力者層は 4 つのテストでいずれも受験者能力が低く、特にスピーキングとライティングにおいて低い傾向が見られた。一方、高い受験者能力者層は他の受験者がリーディングとリスニングでも平均的な受験者能力値が高く推定されることから、相対的にリーディングとリスニングは平均に近づく一方で、スピーキングとライティングが特に高く評価されている傾向が見られた。中間の受験者能力を有するクラスターの学習者は平均的な推定値が得られている。これらの結果から 4 技能を受験する学習者に関しては、受験者能力が低い際には受容技能を得意とするが、受験者能力が高くなると発表技能においても力を発揮できる可能性が示された。

図 6

受験者能力のデンドログラムとクラスターごとの能力分布



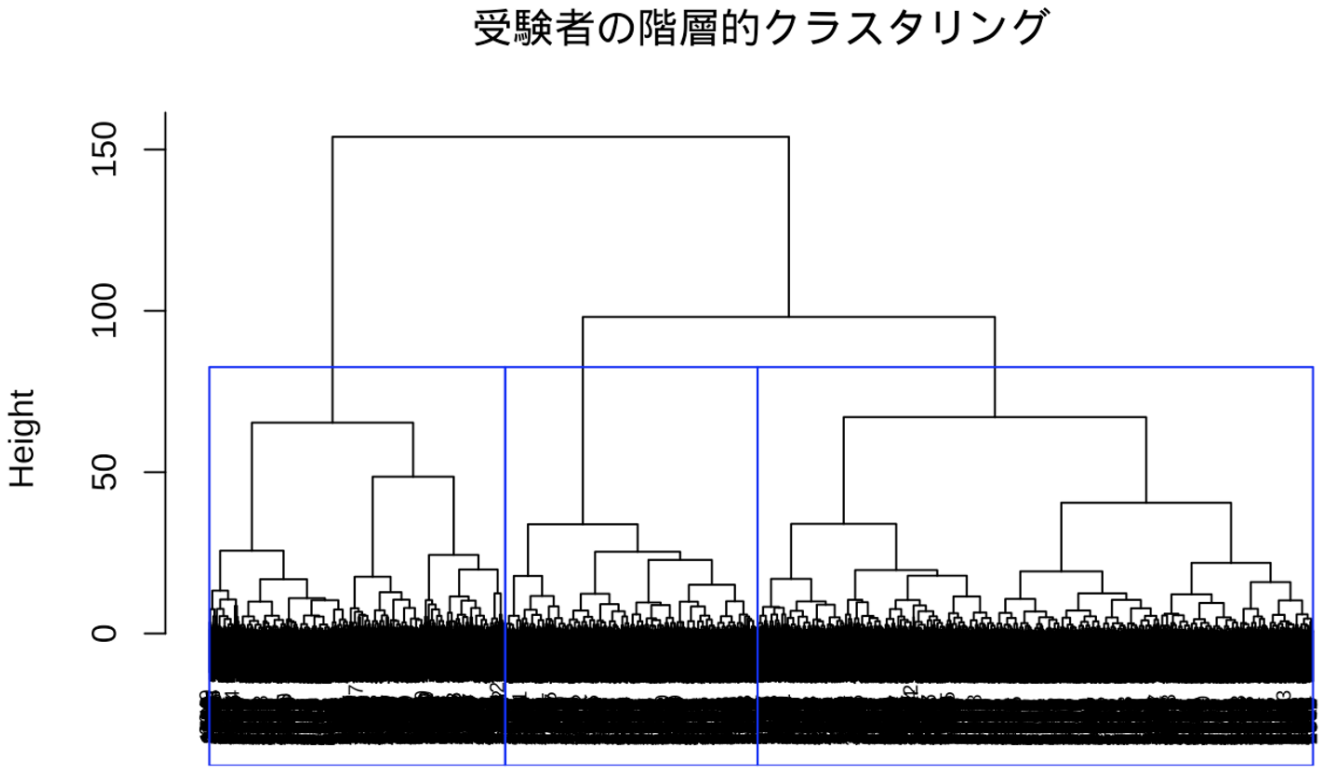
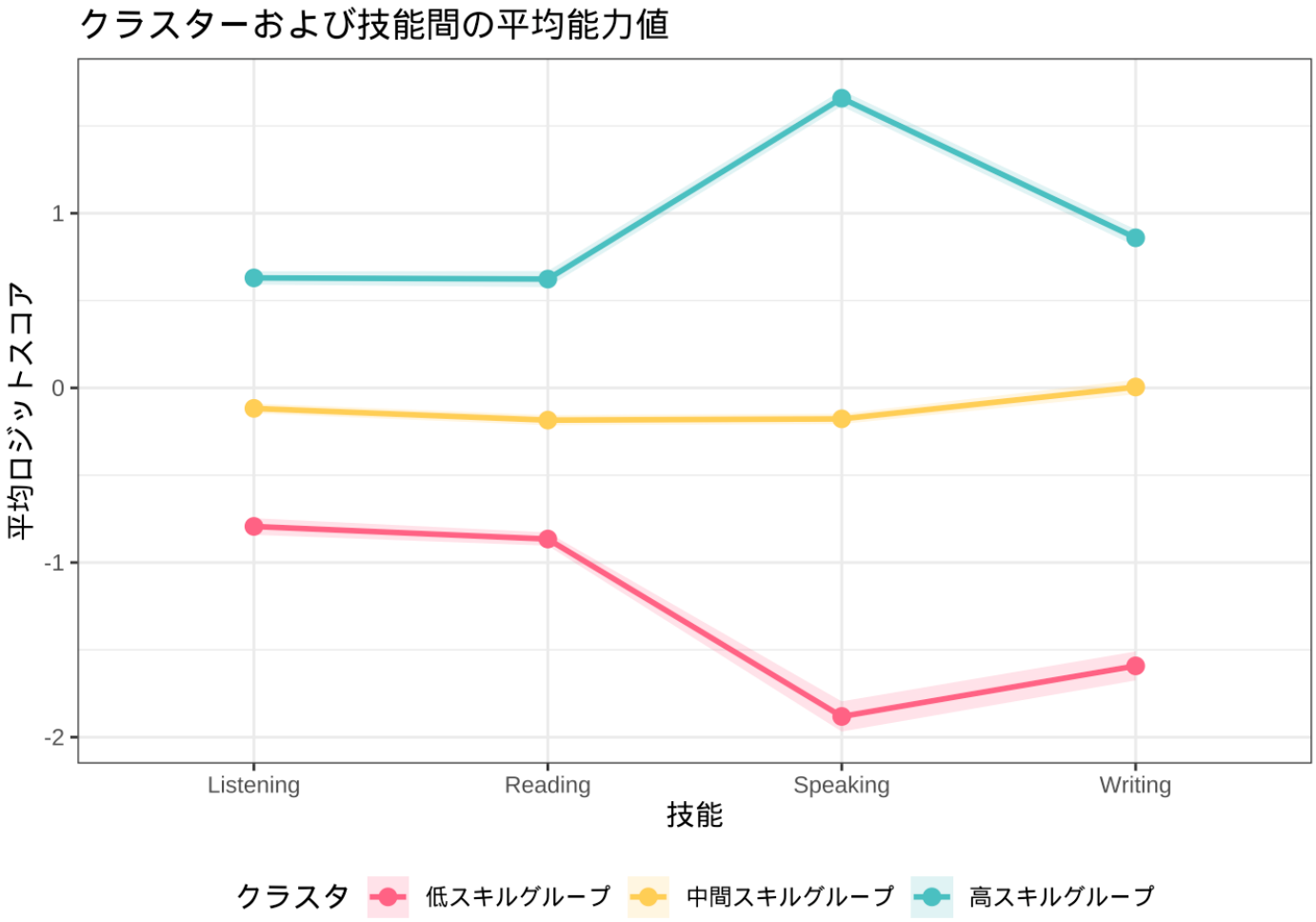


図 7

テストごとのクラスターの受験者能力



クラスターおよび 4 技能における変動係数[5]は表 5 および図 8 に示される通りである。中間クラスター (n = 1634) は受験者数が低クラスター (n = 586) の約 3 倍であり、高クラスター (n = 981) の約 1.5 倍である。また中間クラスターのライティングでは受験者の推定能力値の平均が 0.0036、標準偏差が-0.022 であるため、変動係数は非常に大きな値が得られている。変動係数の値が小さいことはクラスター内のばらつきが小さいことを意味している。

低および高クラスターの変動係数が 1 を下回っており、4 技能間で推定された受験者能力のばらつきは少ない。言い換えると、受験者能力が高いクラスターではリスニング・リーディングよりもスピーキング・ライティングで推定される能力が高く評価されており、反対に受験者能力が低いクラスターではリスニング・リーディングよりもスピーキング・ライティングで推定される能力が低く評価されている。発信技能に自信のある受験者は、L&R で全問正解を目指した正確かつ効率的な英語使用を、発信技能には自信のない受験者は S&W の受験を通して、特にスピーキング能力の伸長を目指す方向性が考えられる。なお、これらの傾向は低クラスターに属する受験者が全体の 18.3%、高クラスターは 30.6%であることを踏まえると、全体的な傾向であるとは断定できないことに注意が必要である。

[5] 標準偏差を平均値の絶対値で割った値。単位の異なるデータセット間や、平均値が大きく異なるグループ間でのばらつきを比較できる。

表 5

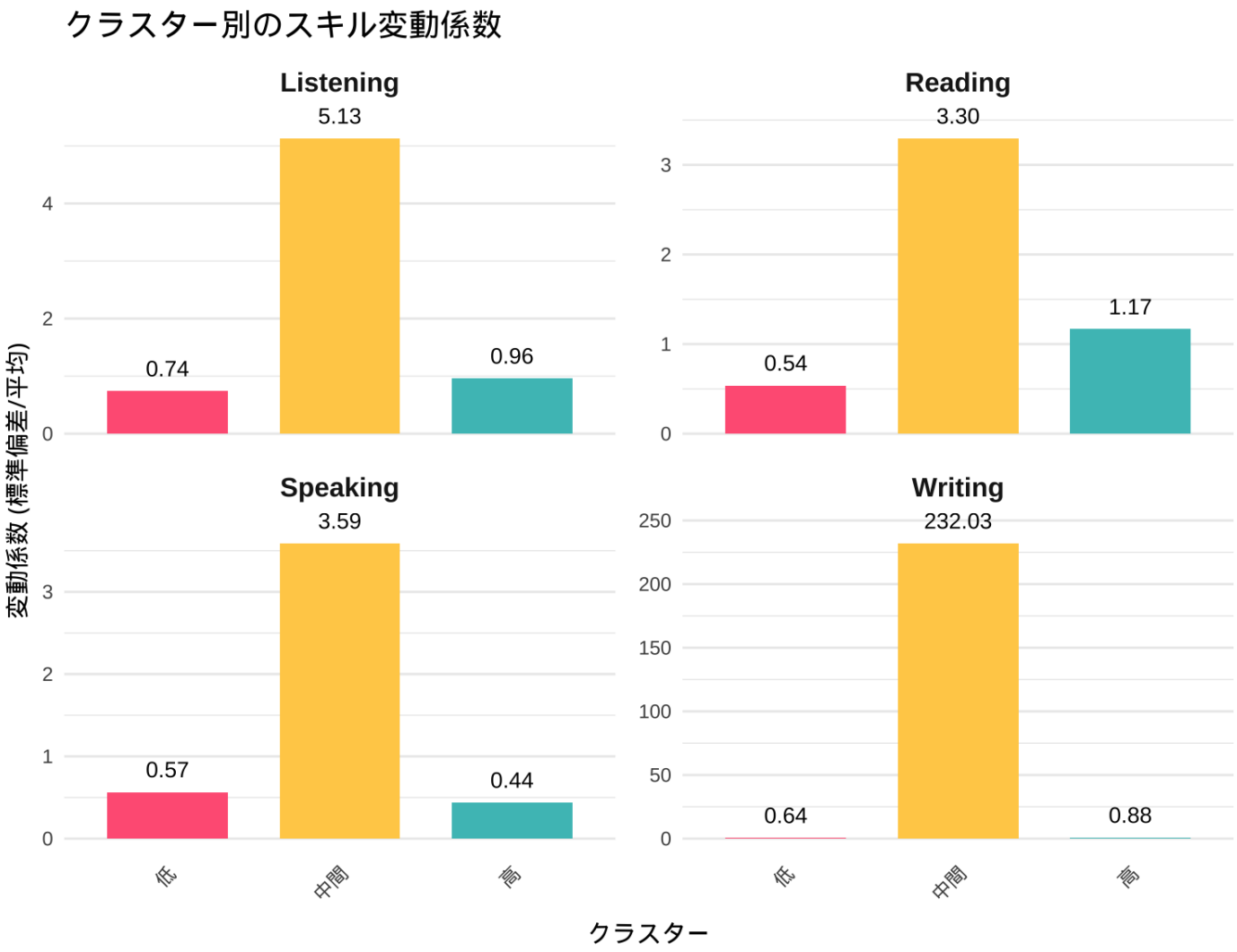
クラスターおよび 4 技能における変動係数

クラスター	受験者数	リスニング	リーディング	スピーキング	ライティング
中間	1634 (51.0%)	5.13	3.30	3.59	232.03
低	586 (18.3%)	0.74	0.54	0.57	0.64
高	981 (30.6%)	0.96	1.17	0.44	0.88

注. 4 技能のテスト全てを受験した受験者データのみを対象としているため受験者数が分析データと異なる。

図 8

クラスターおよび 4 技能における変動係数の棒グラフ

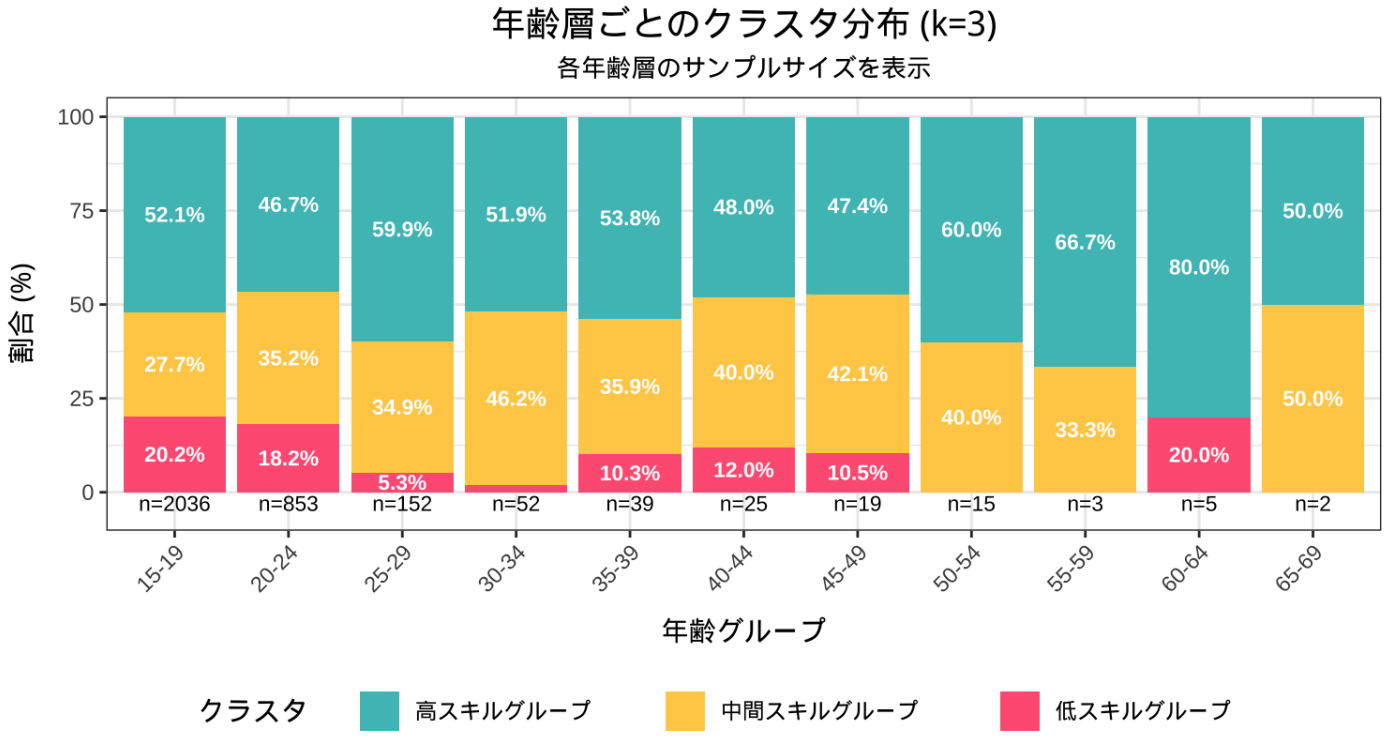


2.6 Differential Person Function：性別と年齢の影響の検討

受験者能力の推定に受験者の特性が与える影響を検討した結果は図 9 の通りである。性別はテストおよびクラスターの間で差は見られなかったが、年齢については異なる傾向が見られた。ただし、今回のデータでは 40 代以上の受験者が少ないことから、明確な判断をつけることは難しい。今後は年齢以外の受験者特性 (e.g., 職業・普段の英語学習・動機づけ) がどのように影響を与えているのかを検討する必要がある。

図 9

Differential Person Function



2.7 TOEIC L&R および TOEIC S&W の結果と考察

本研究では 4 技能テスト受験者のデータを対象に、潜在変数である受験者能力を GLMM によって推定し、その受験者能力に基づいてテストごとの特徴をクラスター分析で検討した。推定された受験者能力は 4 技能の間で差はなかったが (図 3)、中程度から大程度の正の相関が見られた (図 4)。具体的には、リーディングとリスニングの間では r = .80 の相関が見られた、スピーキングとライティングの間では rs = .50–.60 の中程度の相関が見られた。全体的な傾向からは、個別の技能が熟達した学習は他の技能も同じように熟達している傾向が示された。しかし、最低・最高スコアおよび採点スケールの関係性が技能間で異なるように (図 5)、受験者能力にはいくつかの下位区分が存在する可能性が示された。

クラスター分析の結果は主に以下の 2 点である (図 6)。(1) 受験者能力が低い際には、リーディングとリスニングにおける受験者能力が、スピーキングとライティングにおける受験者能力よりも高い。(2) 受験者能力が高くなると、スピーキングとライティングにおいても高い能力を発揮できる。ライティングおよびスピーキング能力における受験者能力のばらつきは、リーディングやリスニングとは統計的には差はなかったが、最低・最高スコアおよび採点スケールを比較した図 5 においては一定のばらつきが見られた。このばらつきは受験者クラスターに応じて異なる傾向が見られたことから、4 技能学習について以下のような示唆が考えられる。(a) リスニングやリーディングでは平均的な得点を得ている学習者であっても、スピーキングとライティングには苦手意識を有

している学習者は約 3 割存在する。このようなケースでは、受容技能で学習するような難易度の高い単語を使わずともタスクパフォーマンスは十分に高めることができたり、写真描写問題には定型表現で回答できたりすることを踏まえ、流暢性の向上や発信技能の自動化を目指すと良いかもしれない。(b) 高い受験者能力を有する学習者の約 2 割はスピーキングとライティングにおいて、リーディングやリスニングよりも高い得点を得ている。このような学習者は L&R においても全問正解を目指し、正確かつ効率的な英語使用を心がけると良いかもしれない。(c) 大部分が属する中間層は 4 技能が平均的に並んでおり、4 技能の能力がバランスよく身についていることが示された (図 7)。また、図 7 の Y 軸には受験者能力が示されており、黄色の中間スキルグループから、青の高いスキルグループへと成長する傾向として、リーディングとリスニングは同時並行的に伸びていき、スピーキングが特に顕著に成長する傾向が見られた。そのため、より英語使用の上級者を目指す際には、リーディングとリスニングの学習は継続しつつ、発表技能であるスピーキングとライティングに力を注ぐことが重要であろう。さらに、中間スキルグループ同士では、ライティングにおける能力のばらつきが非常に大きいことから、周りと差をつけたい学習者はライティングに力を入れた学習をするとよいかもしれない。

なお、これらの結果は性別による影響は見られないが、年齢や他の受験者要因に影響を受けている可能性がある (図 9)。今後は受験者能力の推定に影響を与える要因を検討することも必要である。

3. TOEIC Bridge L&R および TOEIC Bridge S&W の分析

TOEIC Bridge L&R および TOEIC Bridge S&W のデータについても、TOEIC L&W および TOEIC S&W の分析手順と同様に分析を行った。分析対象とするのはオンラインで受験した 2862 名である。なお、TOEIC Bridge についても技能別の受験が可能であるため、欠損パターンは 3 つである。最終的な分析対象となった受験者数は 2487 名である。TOEIC Bridge L&R のテストフォームは 9 件、TOEIC Bridge S&W のテストフォームは 8 件である。

3.1 記述統計

記述統計は表 6・表 7 および図 10 の通りである。素点からは、リスニングの方がリーディングよりも正解率が高く、スピーキングの方がライティングよりも得点が高い傾向が見られる。受容技能に関しては TOEIC L&R と同様の結果である一方、産出技能に関してはスピーキングの方が高いスコアを示しており、TOEIC S&W とは異なる傾向が見られた。

表 6

TOEIC Bridge L&R および S&W の記述統計

	Listening		Reading	
Part	*M*	*SD*	*M*	*SD*
1	82.44	19.26	60.50	18.94
2	66.23	19.97	61.23	20.58
3	71.28	22.14	58.12	23.13
4	50.37	18.45	-	-

注. Listening と Reading は 2 値データであることから正解率を表している。Listening の Part 1 は画像選択問題、Part 2 は応答問題、Part 3 は会話問題、Part 4 は説明文問題。Reading の Part 1 は短文穴埋め問題、Part 2 は長文穴埋め問題、Part 3 は読解問題。

表 7

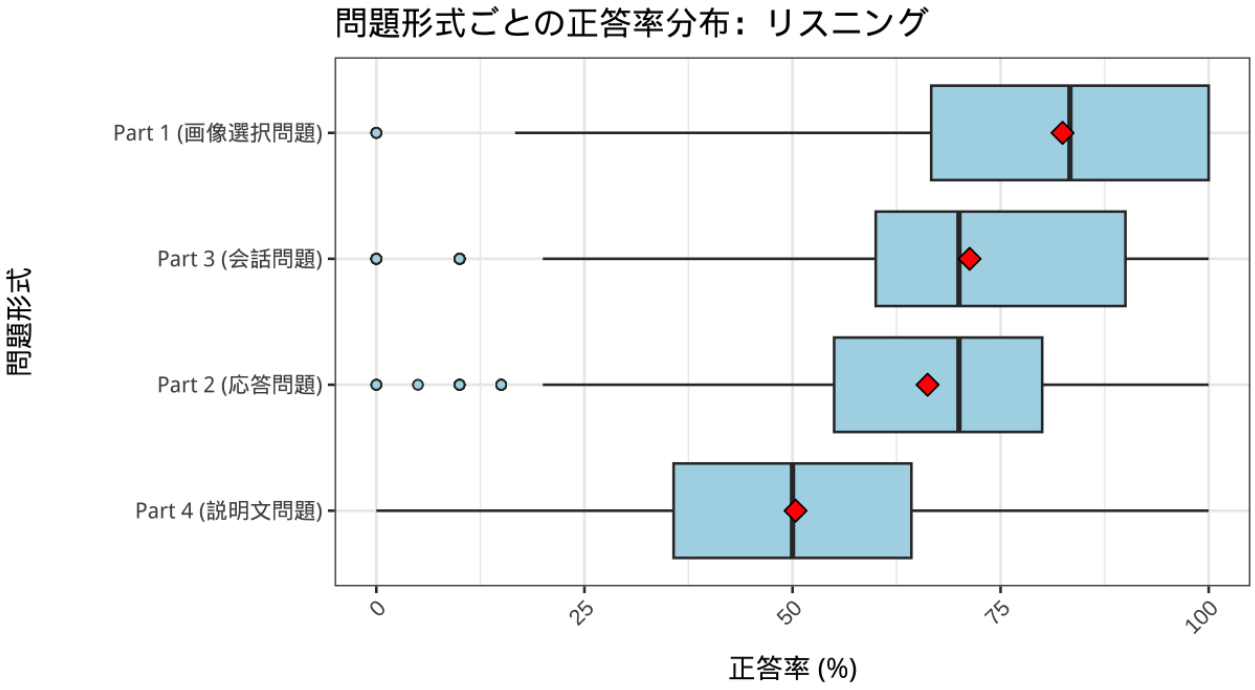
TOEIC Bridge S&W の記述統計

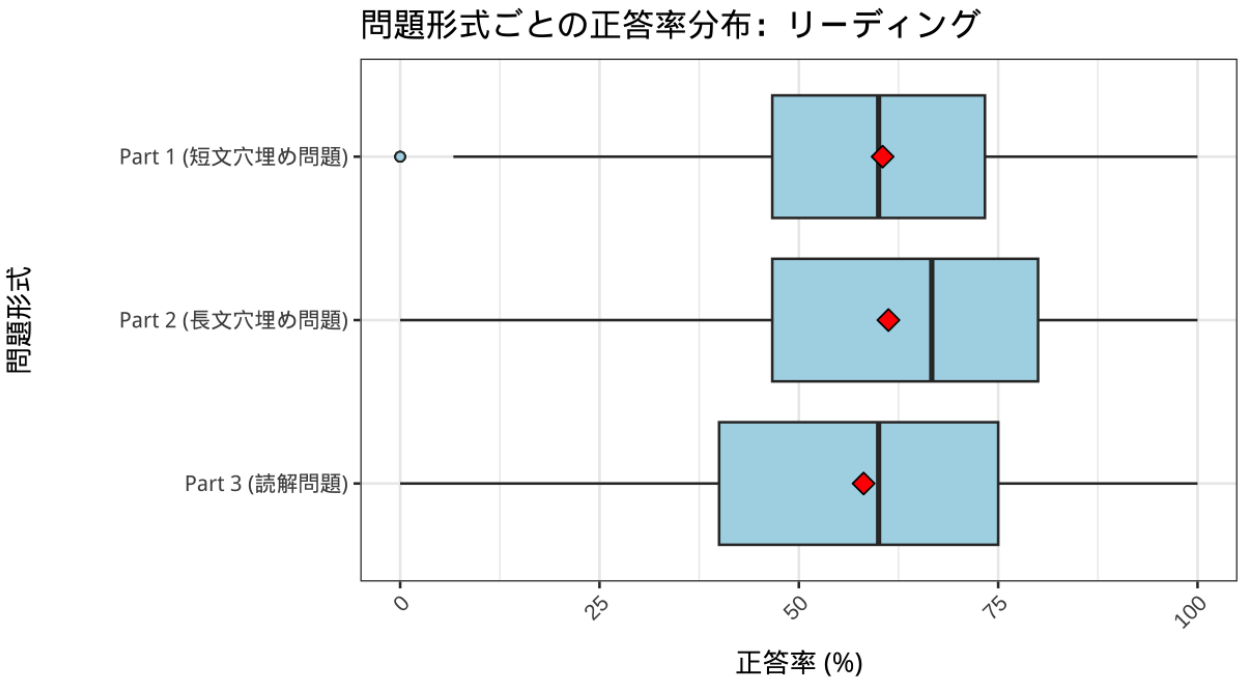
	Writing		Speaking	
Question	*M*	*SD*	*M*	*SD*
1	1.74	0.47	2.39	0.64
2	1.47	0.51	2.35	0.62
3	1.75	0.44	2.02	0.65
4	2.22	0.85	2.16	0.62
5	1.97	0.75	1..49	0.78
6	2.18	0.77	1.71	0.76
7	1.73	0.88	2.44	0.82
8	1.47	0.78	2.29	0.83
9	2.42	0.99	-	-

注. Speaking における QUESTION 1〜2 は音読問題、QUESTION 3〜4 は写真描写問題、QUESTION 5 は聞いたことを伝える問題、QUESTION 6 は短い応答問題、QUESTION 7 はストーリー作成問題、QUESTION 8 はアドバイスをする問題。採点スケールは QUESTION 1〜6 は 0〜3、QUESTION 7〜8 は 0〜4。Writing における QUESTION 1〜3 は文を組み立てる問題、QUESTION 4〜6 は写真描写問題、QUESTION 7 は短文メッセージ返信問題、QUESTION 8 はストーリー記述問題、QUESTION 9 は長文メッセージ返信問題。採点スケールは QUESTION 1〜3 は 0〜2、QUESTION 4〜8 は 0〜3、QUESTION 9 は 0〜4。

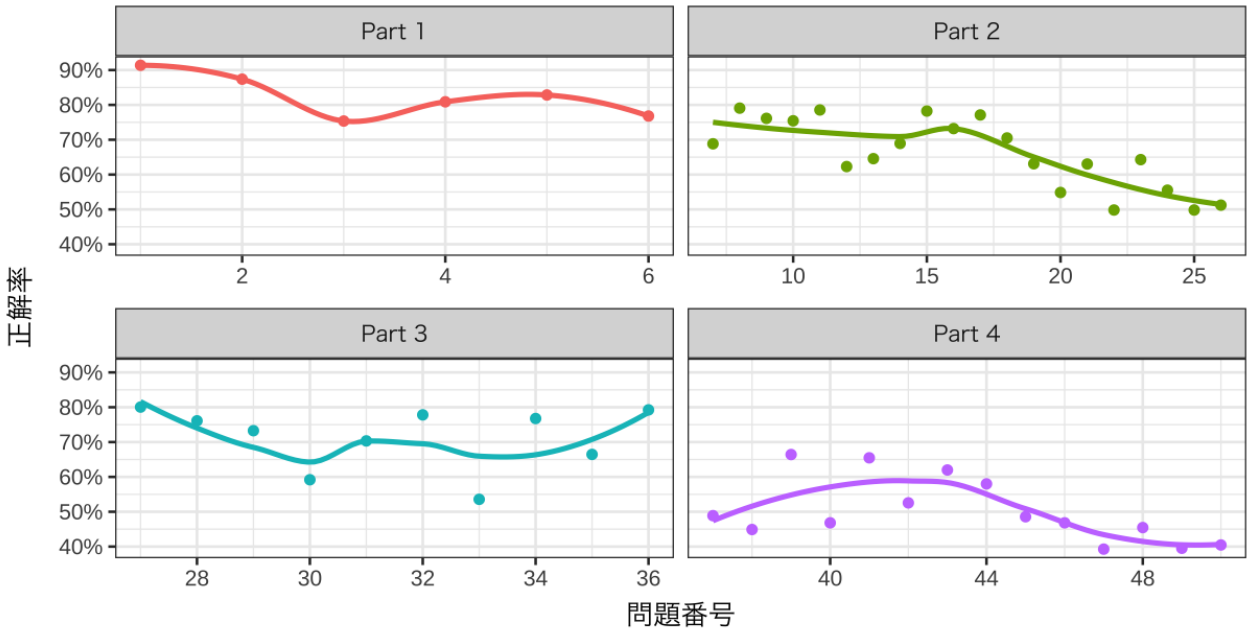
図 10

TOEIC Bridge L&R および Bridge S&W の記述統計

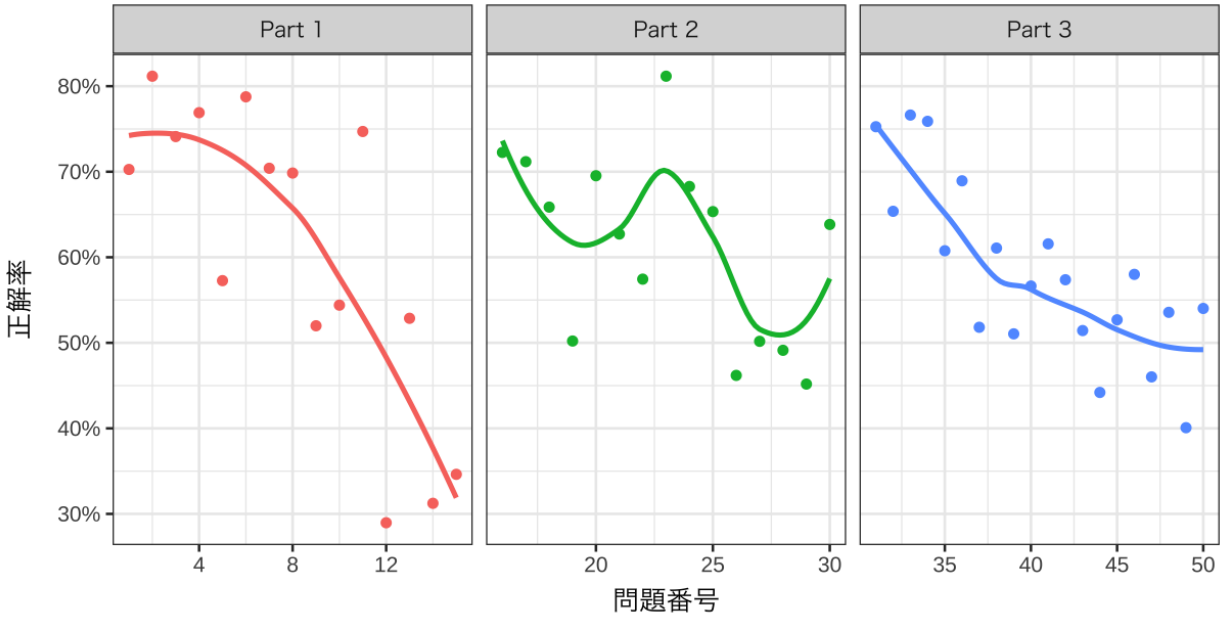




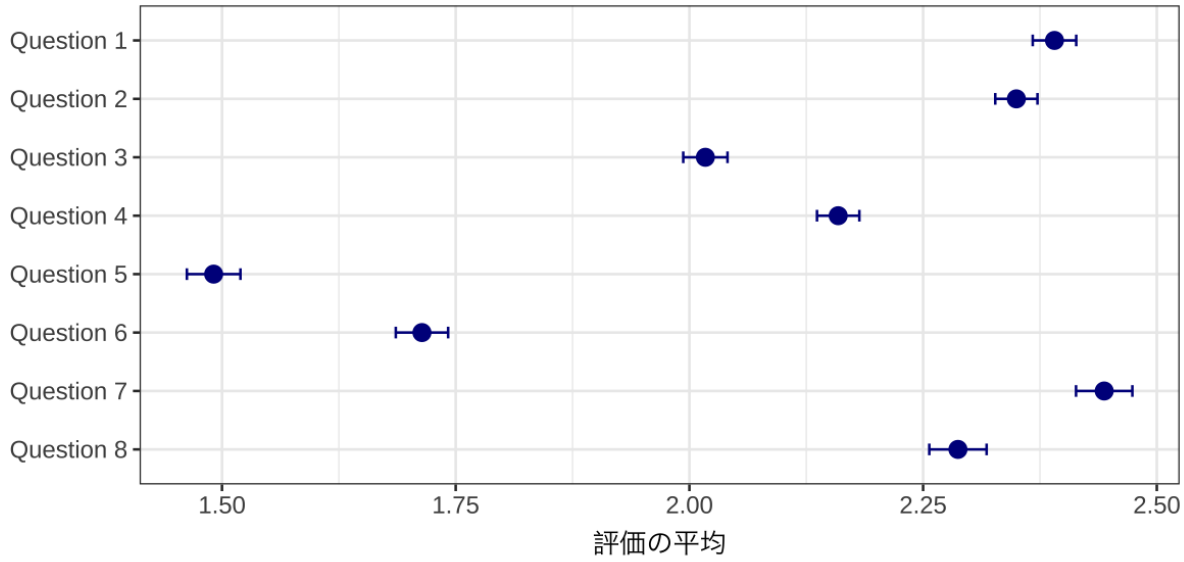
リスニング：問題番号別正解率



リーディング：問題番号別正解率

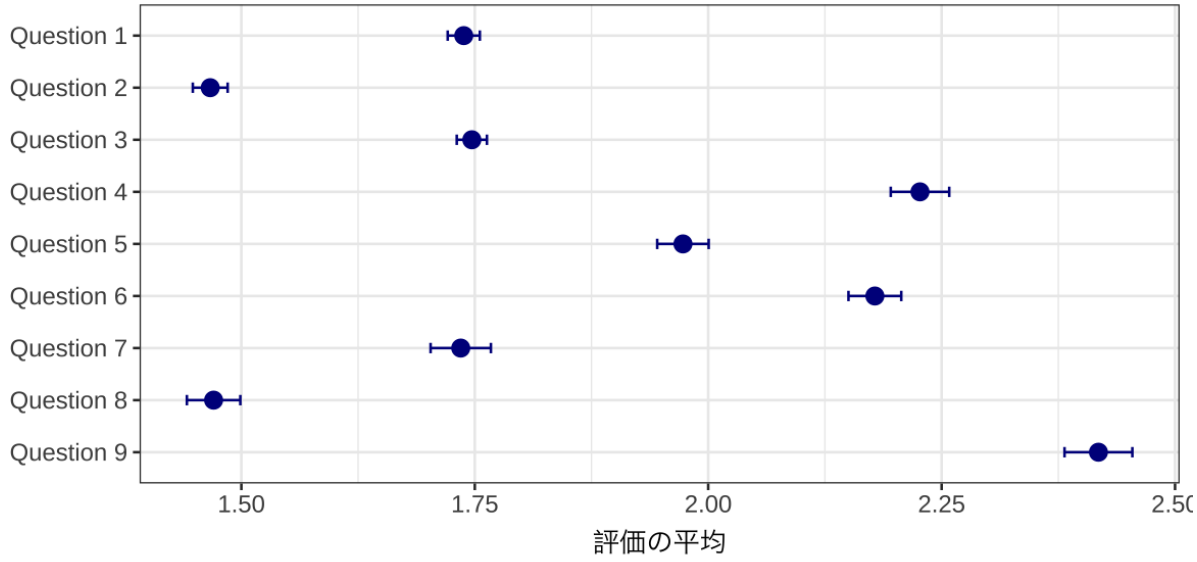


スピーキング：タスク別評価の平均と95%信頼区間



注：エラーバーは95%信頼区間を示す。
採点スケール：Question1-6は0 ~ 3、Question7,8は0 ~ 4

ライティング：タスク別評価の平均と95%信頼区間



注：エラーバーは95%信頼区間を示す。
採点スケール：Question1-3は0 ~ 2、Question4-8は0 ~ 3、Question9は0 ~ 4。

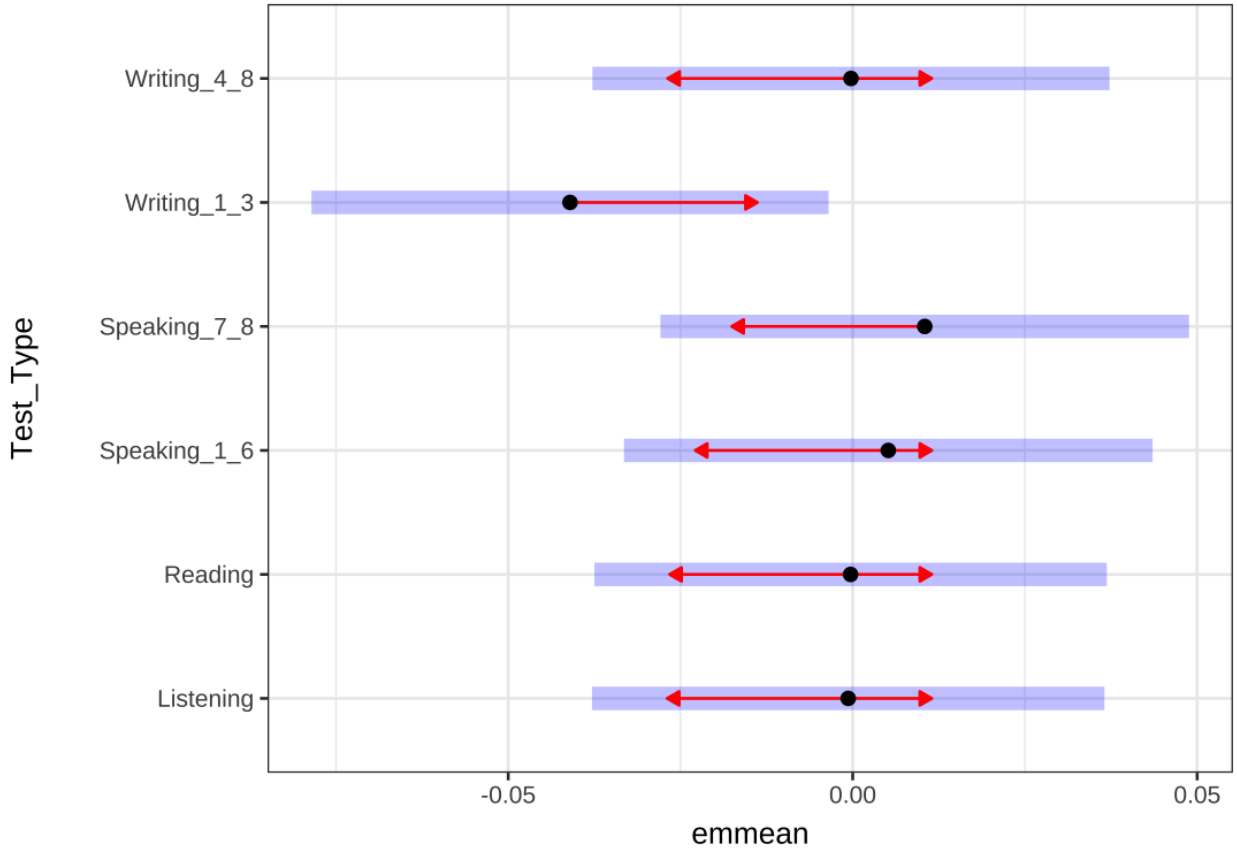
3.2 受験者能力の推定

TOEIC L&R および S&W と同様に、受験者能力を GLMM で推定した。なお、1 水準のみである採点スケール (ライティングの QUESTION 9) は GLMM にて推定できないため、結果は表示されていない[6]。図 11 に基づくと、受験者能力の平均値は 4 技能間でほとんど変わらないようにみられるが、スピーキングにおいて受験者能力の推定値が-4 から+3 をカバーしており、受験者能力が混在していることがわかる。なお、受験者能力値の解釈は値が低いほど、高い受験者能力であることを意味している。

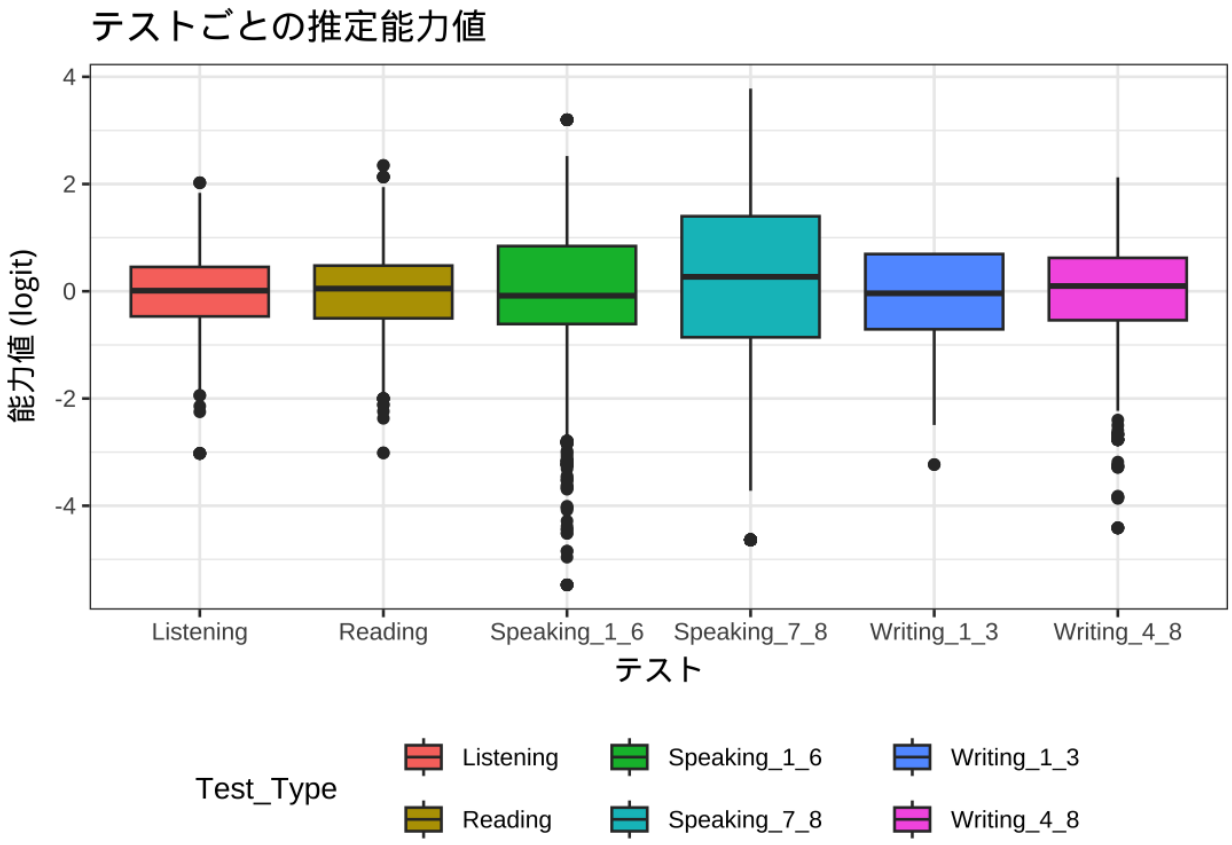
技能間での多重比較の結果、ライティングの QUESTION 1〜3 (文を組み立てる問題) の推定値は他と比べて有意に低かった。つまり、文を組み立てる問題では他と比べて高得点を獲得している傾向があった。それ以外の比較 (例：リスニングとスピーキング、リーディングとライティングの QUESTION 4〜8) の間では有意な差は見られなかった。

図 11

受験者能力の推定値



[6] TOEIC L&R および S&W と同様。


テストごとの推定能力値
4
2
0
-2
-4
能力値 (logit)
Listening
Reading
Speaking_1_6
Speaking_7_8
Writing_1_3
Writing_4_8
テスト
Test_Type
Listening
Reading
Speaking_1_6
Speaking_7_8
Writing_1_3
Writing_4_8

3.3 4 技能間の受験者能力の相関

受験者能力の最低・最高のスコアおよび採点スケールは図 12、相関係数は図 13 に示される通りである。技能間で大きなばらつきが見られるが、中から大の正の相関が示された。これらの結果は TOEIC L&R および S&W と同様である。つまり、TOEIC Bridge において測定される構成概念の内的妥当性は TOEIC L&R・S&W と類似していると考えられる。

図 12

各テスト・タスクの最低・最高のスコアおよび採点スケールの分布

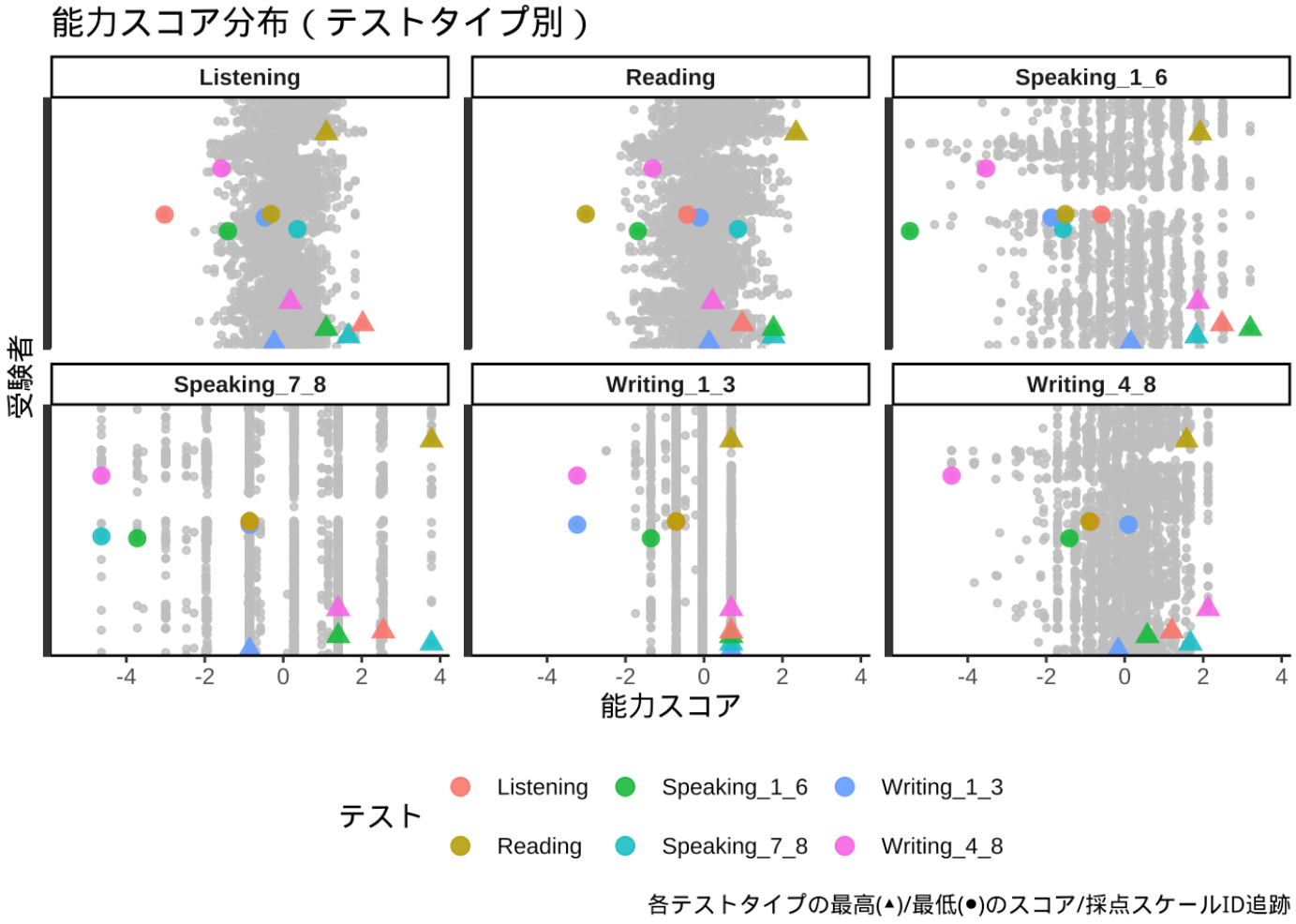
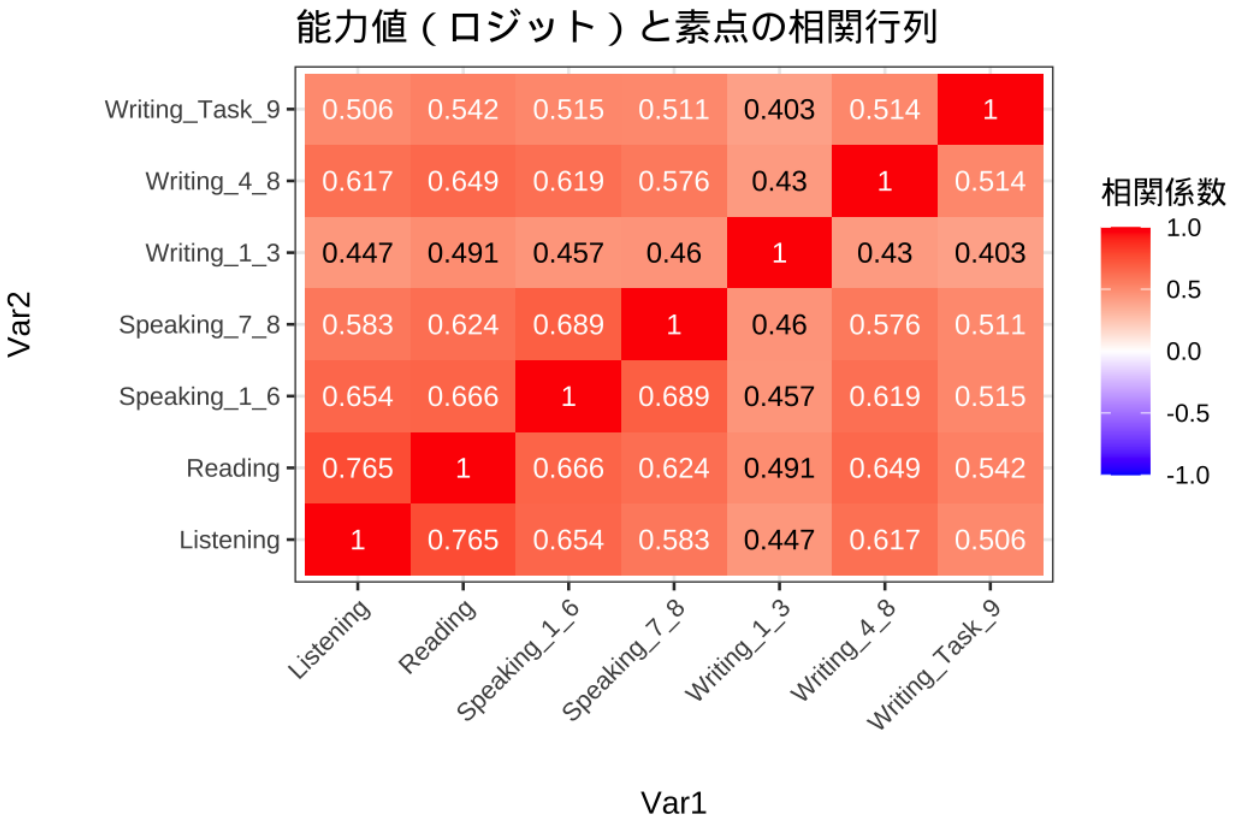
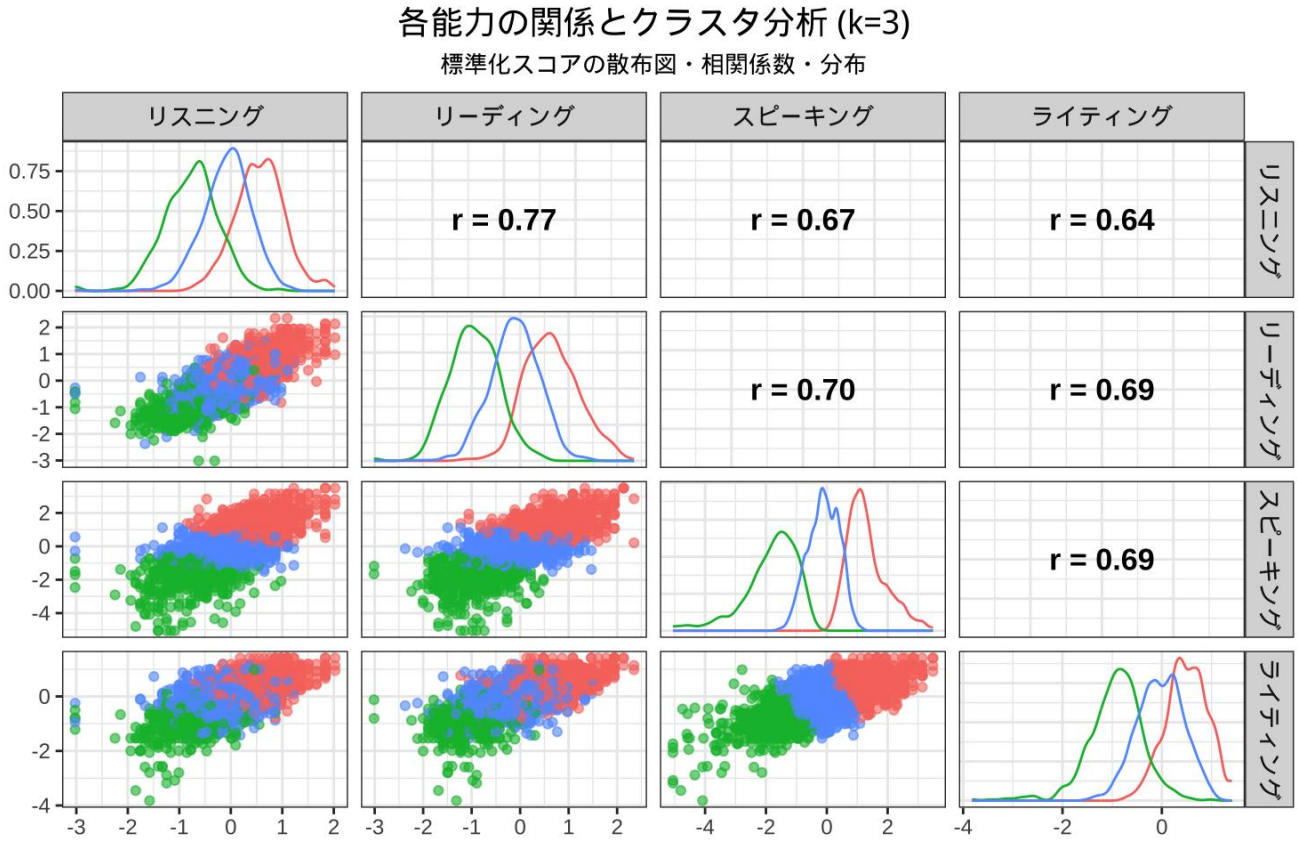


図 13

4 技能間の受験者能力の相関係数





3.4 クラスター分析による能力分布

階層的クラスター分析の結果は図 14 の通りである。TOEIC L&R および S&W と同様に、クラスターは 3 つが検出されており、クラスターごとの能力分布の結果も TOEIC L&R・S&W と同様である。つまり、受験者能力の高いクラスター、平均的な受験者能力のクラスター、受験者能力の低いクラスターの 3 つが検出された。

クラスターおよび 4 技能の推定値を比較した結果、中間グループは 4 技能の間で推定値が安定していたが、低および高グループではスピーキングにおいて推定値にばらつきが見られた。この傾向は TOEIC L&R および S&W と同様である。つまり、初中級学習者を対象とした TOEIC Bridge L&R・S&W においてもスピーキング能力の発達は他の技能とは異なる可能性が示された。

図 14

受験者能力のデンドログラム

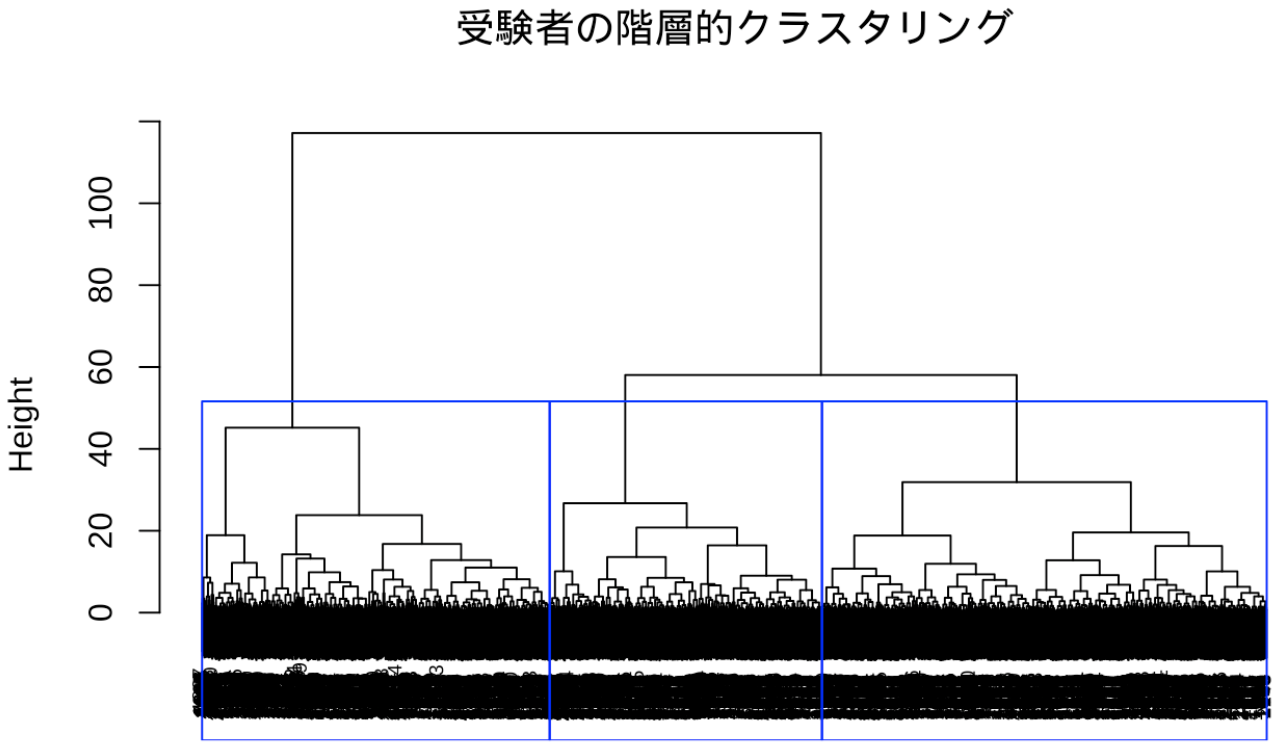
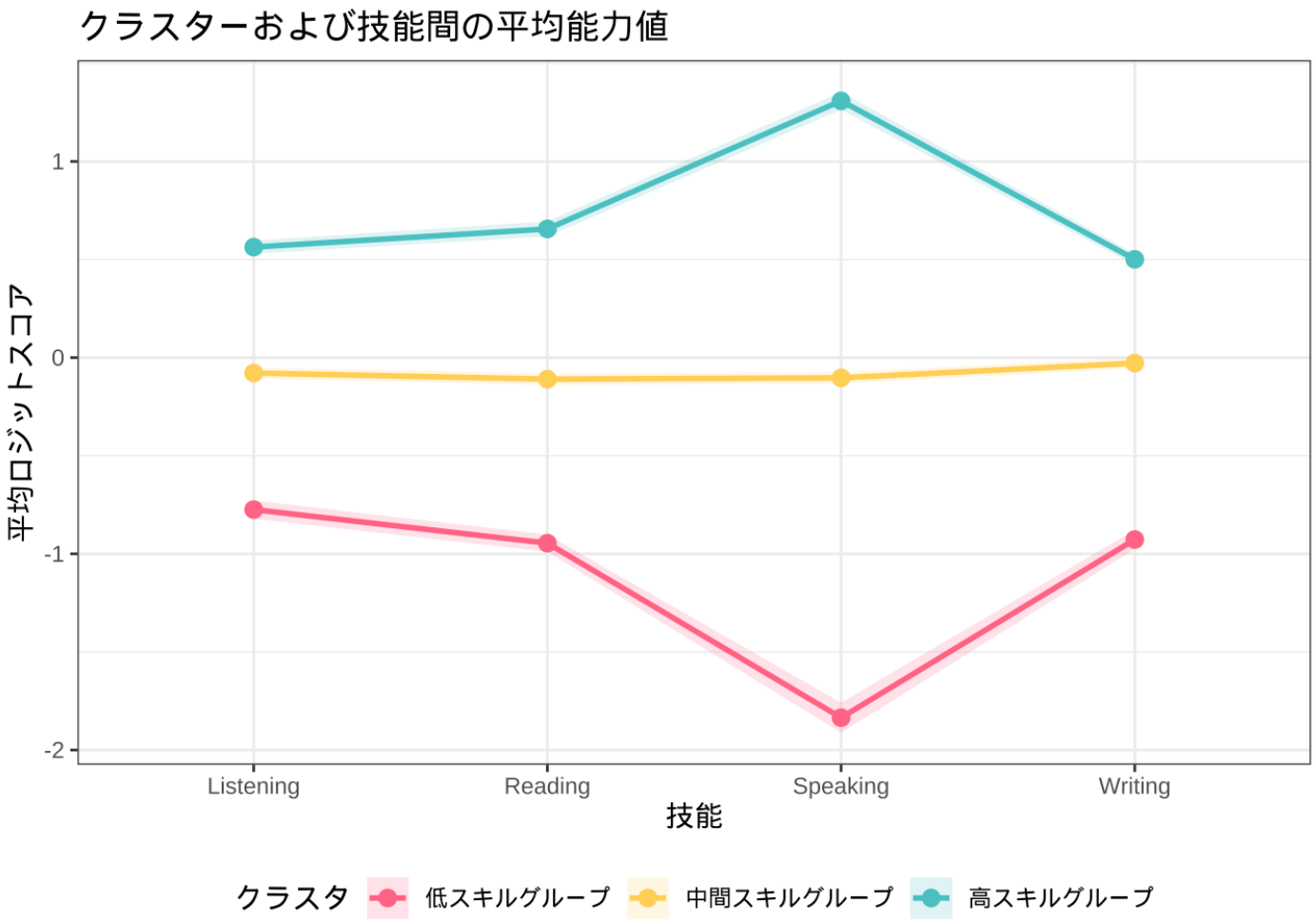


図 15

テストごとのクラスターの能力分布

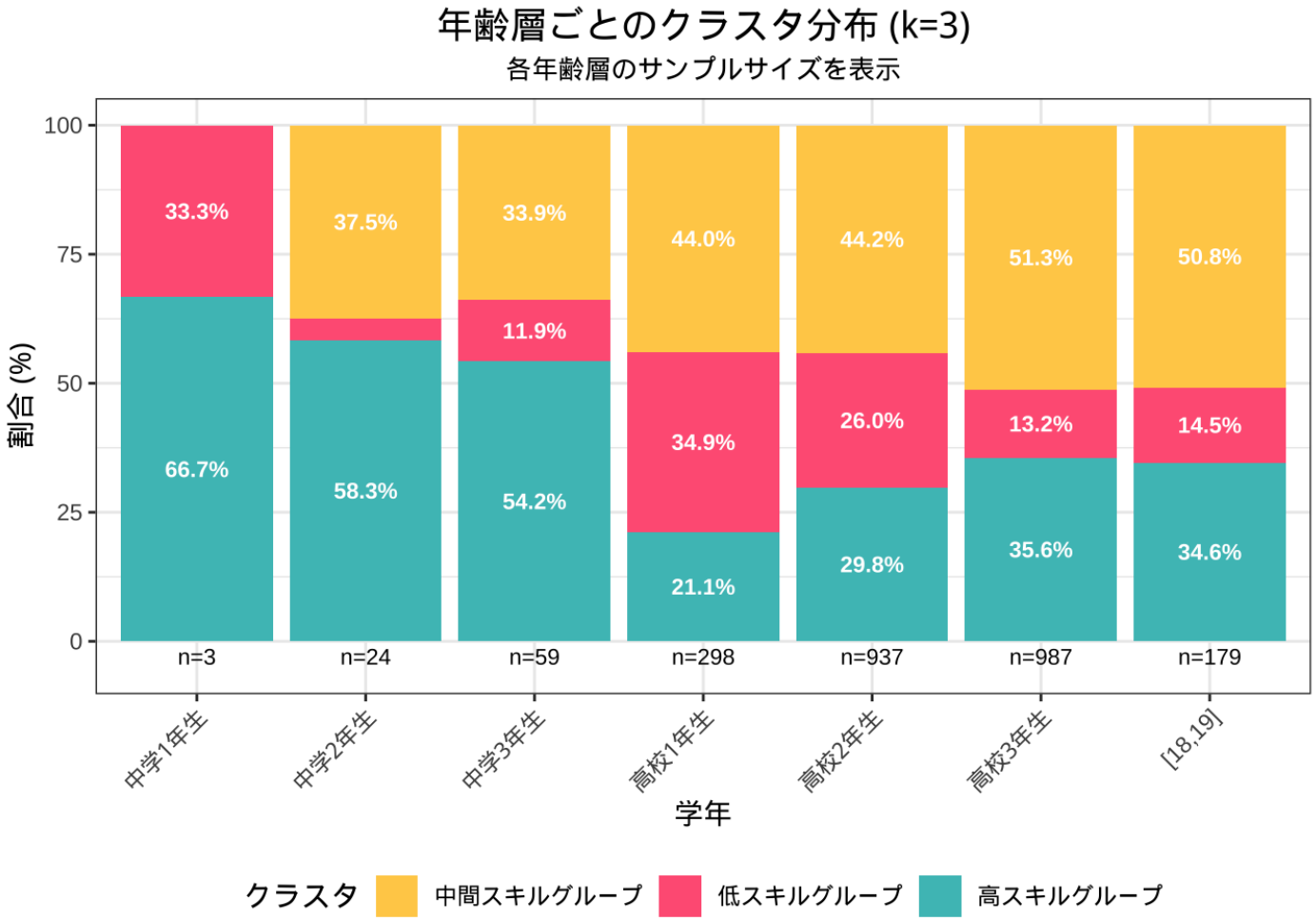


3.5 Differential Person Function：年齢の影響

初中級学習者がターゲットとなっている TOEIC Bridge は若年層の受験も多く、結果に影響を与える受験者要因として、受験者の年齢が挙げられる。なお、今回のデータは若年層に偏っていたことに注意が必要である。図 16 は年齢層ごとのクラスターの割合を可視化したものである。10〜12 歳の受験者は 3 名であることから解釈を控えるが、中学生・高校生を中心とする受験者では、年齢が高くなるにつれて低受験者の割合が増加する傾向が見られた。また、13〜14 歳の受験者 (中学 1〜2 年生) の半数は中間スキルグループに属しているが、15〜16 歳 (中学 3 年〜高校 1 年) となると、低・高スキルグループの割合が増えていた。さらに、17〜19 歳では高スキルグループに所属する学習者は 3 割であり、半数が中間層に属していた。

図 16

Differential Person Function



3.6 TOEIC Bridge L&R および TOEIC Bridge S&W の結果と考察

TOEIC Bridge L&R および TOEIC Bridge S&W における受験者能力の推定は、受験者能力の相関分析において中程度の正の相関が見られ、受験者能力のクラスター分析においても平均的な能力の受験者、受験者能力の高い層、低い層の 3 つのクラスターが見られた。また、スピーキング能力の推定値がクラスター間で異なり、全体的にライティングにおける文を組み立てる問題は他のタスク・問題よりも高い得点を得ている傾向が見られた。

TOEIC Bridge に関する分析の結果から、4 技能学習について以下のような示唆が考えられる。GLMM の下位検定の結果、文を組み立てる問題では受験者全体が他の問題よりも有意に異なる推定受験者能力を示し (図 11)、受験者能力のクラスターはスピーキングで大きく差が開いていたことから、(a) 受験者は発表技能としての文法知識を十分に身につけているが、(b) パフォーマンスの中で自動化された文法知識として身についているとは言い難い。リーディングとライティングの相関が中程度の正の値 ($r = .64$) であったことから、受容技能として明示的に学習した文法知識をコミュニケーションタスクの中で使いこなすために、練習の機会を繰り返し設けることが重要である。中学校 1・2 年生の段階では約半数が所属していた中間スキルグループから、中学校 3 年生・高校 1 年生の段階で低スキルグループへと移ってしまうことを避けるためにも、教室環境においては目的・場面・状況を柔軟に設定したコミュニケーション課題が重要である。なお、対象データ規模が TOEIC L&R・S&W と約 1000 名と異なるにもかかわらず、スピーキングの推定能力値の変動を含め、4 技能間で同様の傾向を示した。このことから、学習が初期段階であったとしても (a) 英語に苦手意識のある学習者は話すことから始め、(b) 発信技能に自信がついたらリーディング・リスニングの正確性を高めることが効果的である。

4. 総合考察

本研究は TOEIC Program の各テストを開発、制作する受験結果データの提供を受け、TOEIC L&R・TOEIC S&W・TOEIC Bridge L&R・TOEIC Bridge S&W における受験結果について、同一尺度の受験者能力を用いて 4 技能を比較した。なお、その分析手法の検討及び分析はすべて本著者によるものであり、本調査に ETS の関与はないことを記しておく。

TOEIC L&R・TOEIC S&W (以降まとめて TOEIC テスト) と TOEIC Bridge L&R・TOEIC Bridge S&W (以降まとめて TOEIC Bridge テスト) は問題形式や問題数が大きく異なる。しかし分析によって得られた結果から、(a) 4 技能間には正の相関係数が見られ、(b) 受験者能力によって 3 つのクラスターが形成されていた。正の相関係数から、リーディングの得点が高い学習者はライティングでも高い得点を得ることができるなど、4 技能は連動して発達することが示された。ただし、全体的な能力が低い段階においてはスピーキングが他の 3 つに比べて低い得点となっている傾向があった。また、初中級層を対象とした TOEIC Bridge テストにおいては、年齢の要因によってクラスターの割合が変化した。具体的には、高校年生以上においては高スキルグループに属する受験者が 5 割未満へと減少していた。年齢に関する影響は特に若年層において顕著であるが、その理由として年齢とともに英語に対する苦手意識が増加している可能性や、中学段階の受験者は早期から英語学習を開始している可能性などテスト受験までの学習経験の影響も考えられる。今後は若年層の受験者の特徴がどのような影響を与えるのか、更なる検討が必要である。

一般的に、リスニング・リーディングの受容技能にはある程度自信があるが、スピーキング・ライティングの発表技能については苦手意識を抱える学習者が多い。本研究では 4 技能の間には中程度の正の相関が見られたことから、「英語力」という 4 技能全体的に影響する共通因子が存在すると仮定される。この共通因子に基づくと、受容技能がある程度得意である学習者は、文法や語彙といったスピーキング・ライティングの基礎的な能力は確実に存在していると言える。スピーキング能力をさらに伸長させるには、例えば流暢に話す工夫として、you know

や you mean などのつなぎ言葉を使うことを意識して、発話における自動化を目指すことが挙げられる。

TOEIC テストと TOEIC Bridge テストはともに類似した結果を示しているが、用いられているタスクの違いや受験者集団の違いがあることに注意が必要である。例えば、テストが対象としている受験者の違いによって、タスク達成の難易度や解答に必要となる認知活動の違いが挙げられる。また、年齢が高くなるにつれ、説得的な発話内容の形成や繋ぎ言葉の意図的な方略使用が増え、スピーキングの評価観点である内容や流暢性が高まる可能性もある。TOEIC テストは実社会における英語使用を想定していることから、飛行機に乗る経験や E メールの経験など、現実世界における背景知識を最大限に活用することも必要となる可能性がある。さらに、受験回数によって、特定のストラテジーを取る受験生の存在が想定される。具体的には、写真描写課題では I can see や There are などの決まったパターンの使用、ライティングでは賛成意見と反対意見のメリット・デメリットを天秤にかけ、自分の主張がデメリットを上回ることを主張するなど、論理の組み立てなどのストラテジーが存在する。そのため、受験回数などの繰り返し練習で異なる結果が得られる可能性がある。今後はさらに詳細な問題項目の分析や受験者特性を踏まえた分析が必要となる。

5. 結論

本研究は TOEIC テストと TOEIC Bridge テストにおける 4 技能の関係性について、共通受験者デザインを用いて検討した。得られた結果から、4 つの技能の間には中程度の正の相関があることから「英語能力」を多角的かつ総合的に測定しており、受験者能力は上・中・下の 3 段階に弁別されうる可能性が示された。ただし、この受験者能力の段階別で発表技能の間には差が見られ、特にスピーキング能力の発達が大きく異なることが示された。具体的には、受験者能力が低い際にはスピーキング能力が他と比べて特に低いが、能力が高い受験者はスピーキング能力が特に高く推定される。

TOEIC L&R および TOEIC S&W では、平均的な受験者能力層 (今回のデータではリーディングで 263.28 点、リスニングで 331.45 点、スピーキングで 109.67 点、ライティングで 126.42 点) の場合には、L&R だけではなく S&W を含めた 4 技能を受験することで、自分の得意・苦手な技能が何かを明らかにできる。つまり、TOEIC L&R で 500 点から 700 点の受験者は TOEIC S&W を診断テストとして受験し、スコアレポートを活用しながら、自分の得意な技能をさらに磨く、不得手な技能を練習することなどが考えられる。TOEIC L&R で 700 点を超える中級から上級学習者は、スピーキング能力はもちろん、論理構成や説得的なライティング能力をさらに磨くことが奨励される。

今回の調査で用いたデータに基づくと、低受験者能力のクラスターに該当する学習者は 4 技能の中でも特にスピーキングにおいて相対的に低い受験者能力であると推定されていた。4 技能の推定能力値の間では中程度以上の正の相関が確認されており、平均的な能力クラスターへと英語技能が上昇する際には受容よりも発信の側面が特に上昇すると想定される。つまり、L&R の受容能力が上昇するにつれてスピーキング能力も同時に上昇することが想定されるが、中程度の熟達度層へと一気にステップアップするためには特に発信能力を磨くことが全体的な英語技能を伸ばすための近道となる。これはアウトプットを行うことによって、自分が英語を使って何ができて何ができないのかをメタ認知できるためである。方略を駆使した英語学習では、自分がスピーキングで使えない単語がある場合には別の簡単な単語を使うパラフレーズが挙げられる。例えば、写真描写課題において「海でカモメが飛んでいる」と言いたいが、「カモメ」を指す英単語がわからない場合には上位語である bird を使い、“Birds are flying.” と表現できる。さらに、明示的な文法知識に自信がある場合には、“There is a bird flying above the sea” のように、there 構文と現在分詞の後置修飾を使って表現できる。

4 技能の間で中程度の正の相関が見られたことから、高い発信能力備える受験者は受容能力も相応の力があると考えられる。しかしリスニング・リーディングでは、自分の知らない単語がなくなることはない。そのため、

前後の文脈から推測する能力を磨くことが求められる。さらに既知の単語であっても、前後の文脈によっては異なる意味として用いられることもある。発信能力の高い受験者クラスターの学習者は、リスニング・リーディングで使われる語彙の広さと深さに磨きをかけ、4 技能をバランスよく学習することが必要である。

TOEIC Bridge L&R および TOEIC Bridge S&W では、平均的な受験者能力層 (リスニングで 29.49 点、リーディングで 33.50 点、スピーキングで 32.2 点、ライティングで 37.40 点) との比較に加え、学習環境を踏まえて今後の学習の道筋を計画する必要があるだろう。例えば、10 代でスピーキングに自信がないと感じている学習者は学内のパフォーマンステストの対策を通して、計画的に方略を学ぶこともできる。TOEIC Bridge S&W のタスクから例を挙げると、写真描写タスクでは “I can see XX” や “There is/are XX” などの定型表現、“Let me see.” や “Well…”、“You know” などの繋ぎ言葉を用いることで流暢性を上げるなどの工夫をとることができる。大学入試を控えた高校生や留学を志す大学生はスピーキングだけでなく、リスニングやライティングなど、4 技能を満遍なく学習することが必要である。そのためには、自分の英語能力を客観的に測定する機会を設けることの重要性が再認識され、またその結果を活用した学習が必要となるだろう。

4 技能の統合的・総合的な学習および指導の重要性が指摘されて久しいが、AI を活用した英語使用を念頭に置いたグローバル社会において、英語学習者は今後ますます 4 技能のバランスが求められると考えられる。本調査では 2 つのテストを活用して現在の学習者の状況を分析し、4 技能における発達傾向を分析した。英語学習者は 4 技能のテスト結果を活用することで、より効果的に学習者が自身の習得状況を理解し、今後の学習に活用していくことが期待される。

6. 参考文献

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Brysbaert, M., & Debeer, D. (2025). How to run linear mixed effects analysis for pairwise comparisons? A tutorial and a proposal for the calculation of standardized effect sizes. *Journal of Cognition, 8*(1), 5, 1–36. https://doi.org/10.5334/joc.409

Bürkner, P.-C. (2020). Bayesian item response modeling in R with brms and Stan. *The R Journal, 12*(1), 405–420. https://doi.org/10.32614/RJ-2020-023

Christensen, R. H. B. (2019). *Cumulative link models for ordinal regression with the R package ordinal*. Technical University of Denmark & Christensen Statistics. https://www.jstatsoft.org/

Dunn, K. J. (2024). Random-item Rasch models and explanatory extensions: A worked example using L2 vocabulary test item responses. *Research Methods in Applied Linguistics, 3*, 100143. https://doi.org/10.1016/j.rmal.2024.100143

Gries, S. T. (2021). *Statistics for Linguistics with R: A practical introduction* (3rd revised & extended ed.). De Gruyter Mouton.

Liao, C.-W., Qu, Y., & Morgan, R. (2010). The relationships of test scores measured by the TOEIC® Listening and Reading Test and TOEIC® Speaking and Writing Tests. In D. E. Powers (Ed.), *TOEIC® compendium* (1st ed., pp. 13.1–13.15). Educational Testing Service.

Linck, J. A., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning, 65*(S1), 185–207. https://doi.org/10.1111/lang.12117

Liu, J., & Costanzo, K. (2013). *The relationship among TOEIC® listening, reading, speaking, and writing skills* (TOEIC Compendium 22.1). Educational Testing Service.

日本語版発行日：2025 年 10 月
日本語発行：一般財団法人 国際ビジネスコミュニケーション協会
（The Institute for International Business Communication; IIBC）
〒164-0001　東京都中野区中野 4-10-2　中野セントラルパークサウス 5F

公式サイト　https://www.iibc-global.org