高崎経済大学論集　第50巻　第３・４合併号　2008　125頁～134頁

古典的項目分析とラッシュモデリングを用いた英語熟達度テストの分析

阿　部　真理子　ブライアン・ウィスナー
酒　井　英　樹

Analyzing an English Language Proficiency Test Using Classical Item Analysis and Rasch Modeling

Abe Mariko・Brian Wistner・Sakai Hideki

Summary

This paper examines the insights that can be obtained from applying two methods of test analysis, classical item analysis and Rasch modeling, and also explores if the listening section of the Michigan English Placement Test (MEPT) is appropriate in difficulty for measuring the participants in this research. The participants were Japanese university students who were divided into novice, intermediate, and advanced proficiency levels based on their English entrance examination scores. The data for this study came from the advanced level students. The results of the classical item and Rasch analyses indicated that the difficulty level of the MEPT listening section was appropriate for the sample. Although some items did not discriminate well, all of the test items fit the Rasch model and exhibited low levels of measurement error. The results point toward a need for increasing the number of items on the listening section of the MEPT.

１．はじめに

本研究の目的は、古典的項目分析手法とラッシュ分析手法を用いて、英語熟達度テストの１つであるミシガンテストが、本研究の協力者である日本人英語学習者の熟達度を測定するのに適切であるかどうかについて検証することである。これら２つの分析方法は、これまで SLEP 熟達度テストや Quick Placement Test-Pen and Paper Test（QPT-PPT）などの英語熟達度テストの分析に用いられている（e.g., Brown, 1989; Culligan＆Gorsuch, 1999; Gorsuch＆Culligan, 2000; Fulcher, 1997; Westrick, 2005）。本研究は、ミシガンテスト（Michigan English Placement Test, Corrigan, Dobson, Kellman, Spaan,＆Tyma, 1993）を取り上げて分析するものである。熟達度の測定にふさわしくないテスト項目を特定することができれば、今後ミシガンテストをより有効に利用することが

できるはずである。今回は、ミシガンテストのうち、リスニングスキルを測定するセクション（20項目）を分析対象とする。

２．方法

２．１．参加者

本研究の参加者は､中学高校で６年間英語を学習してきた日本語を母語とする英語学習者である。ミシガンテストのリスニングセクションを受験した者は、98名であった。同一大学の同じ学部に所属する１年生23名（男性17名・女性６名）、２年生74名（男性54名・女性20名）、３年生１名（女性１名）である。参加者は、大学入学試験の英語の得点によって、初級・中級・上級と分けられた英語必修科目のうち、上級レベルの４クラスに属する英語学習者であった。そのため、熟達度のばらつきが比較的小さい集団であったと考えられる。

２．２．手順

ミシガンテストは2007年度の新学期が始まってから、２～３週後に実施された。テストを受けるにあたって、各クラスにおいてすべて同じ受験の手順と注意事項に関する説明が行われた。リスニングテストは、テープ音声が良く聞こえる環境を確保するため、受験に適した教室において実施された。

２．３．測定具

ミシガンテストは、ミシンガン大学英語研修センターにおいて開発された試験であり、1972年から1987年に英語研修センターが閉鎖されるまでの期間､習熟度別クラスを形成するのに利用された。ミシガンテストには同形式の試験問題が３セット（フォームＡからフォームＣ）用意されているが、本研究ではフォームＣを利用した。テストはリスニング20題、文法30題、語彙30題、読解20題の計100題から構成され、リスニングの時間は約５分で、文法・語彙・読解の解答時間は50分である。ミシガン大学英語研修センターにおいてはテストの結果によって、Beginner（０～29点)、Beginner-high（30～47点)、Intermediate-low（48～60点)、Intermediate（61～47点)、Advanced-low（75～84点)、Advanced（85～100点）の習熟度クラスに分けられる。

本研究の分析対象であるミシガンテストのリスニングセクションは、２つのタイプの問題形式に分けられる。１つ目のタイプは、テープの音声が短い質問文を読み上げるのに対して、その質問の答えとして適切なものを３つの選択肢から選ぶものである。２つ目の特徴は、同じく短い文を聞かされて、その内容と合致するものを３つの選択肢から選ぶものである。リスニングセクションは、問題形式の説明と例題と、問題によって構成されている。

２．４．分析方法

本研究では、古典的項目分析とラッシュモデリングの２つの分析手法を用いる。なお、以下の統計用語の訳語は、JLTA Language Testing 用語集編集委員会（2006）に基づく。基礎統計量（平均、標準偏差、95％信頼区間、最小値、最大値、歪度、尖度）や分布図に加えて、集団規準準拠テスト（norm-referenced test）の古典的項目分析として、項目容易度（item facility）と項目弁別力（item discrimination）を計算する（Brown, 2005）。項目容易度は、ある項目の正答率のことである。0.00から1.00までの値をとる。Brown（2005）によれば、適切な項目容易度は.30から.70であるとされている（p.75）。項目弁別力は、ある項目が、テストの合計点の高かった受験者と、テストの合計点の低かった受験者を分ける度合いを示している。具体的には、合計点に基づき、参加者を３等分し、上位群、中位群、下位群を設定する（25％や27％の群を作ることもあるが、本研究ではBrown, 2005にならい、33％を基準とした）。そのうち、ある項目に関して、上位群の正答率から、下位群の正答率を引いたものが、その項目の弁別力となる。項目弁別力については、.40以上がとてもよい項目（very good items）、.30から.39がよい項目であるが改良が必要かもしれない項目（reasonably good, but possibly subject to improvement）、.20から.29は改良が必要な項目（marginal items, usually needing and being subject to improvement）、.19以下はよくない項目で削除したり作り直すことが必要な項目（poor items, to be rejected or improved by revision）とされている（Brown, p.75）。さらに、テストの信頼性係数としてK-R 21、K-R 20を計算し、それぞれの信頼性係数に基づく測定の標準誤差（standard error of measurement）を計算した。信頼性係数は、テスト項目の間に内的一貫性があるかどうかの度合いを示している。例えば、信頼性係数が.91であるとき、分散のうち91％が一貫しており、９％が誤差によるものであると解釈される（Brown, p.175）。測定の標準誤差は、得点の安定性を示している。例えば、100点満点のテストで、測定の標準誤差が５であるとき、ある受験者の得点80点に関して、約68％の確率で真の得点（true score）が75点から85点の間（得点±測定の標準誤差）におさまると解釈される（Brown, p.188）。古典的項目分析には Microsoft Excel 及び SPSS 12.0J for Windows を用いた。

ラッシュモデリングでは、受験者の能力推定値（ability measures）と項目の困難度推定値（difficulty measures）を算出する。推定値は、ロジッツ（logits）と呼ばれる間隔尺度（interval scales）の単位で表される。ロジッツとは、log odds units（オッズの対数の単位）を短くしたものである（靜、2007、p.172）。通常は、-3.0から+3.0程度の範囲で分布する。例えば、能力推定値が1.0ロジッツである学習者が、困難度推定値1.0ロジッツである項目に正答する確率は、50％であり、より低い困難度推定値の項目に正答する確率は高くなり、より高い困難度推定値の項目に正答する確率は低くなる。さらに、データがどの程度ラッシュモデルに適合（fit）しているかどうかという適合度を見る指標として、アウトファット値とインフィット値が計算されるが、本研究では外れ値の影響を抑えたインフィット値を分析する（靜、 p.312）。インフィット値の平方平均は、1.0を中心に分布する。Bond＆Fox（2007）によれば、インフィット値の解釈の基準として、受験者にとっ

て重要な決定をするための多肢選択肢テスト（high stakes multiple-choice tests）の場合、0.8から1.2の範囲を許容範囲としている（p.243）。値が低すぎると、オーバーフィットと呼び、ラッシュモデルに適合しすぎていることを示す。値が高すぎると、アンダーフィットと呼び、ラッシュモデルに適合していないことを示す。ラッシュ分析にはWinSteps 3.63を使用した。

３．結果

３．１．古典的項目分析

表１は、ミシガンテストのリスニングセクションの基礎統計量を示している。このリスニングセクションの最大可能得点は20点である。平均値（9.39点）や、散布度（標準偏差2.24、最小値4.00、最大値16.00）をみると、本研究の参加者に適切なテストであったことが示唆される。また、歪度を標準誤差で割って得られるｚ値は1.21であり、尖度を標準誤差で割って得られるｚ値は0.60であり、これらの絶対値が1.96よりも小さい値であることから、正規分布を逸脱するものでないことが示されている（Field, 2005, p.72）。図１は、ミシガン・リスニングテストの得点分布を示している。ほぼ正規分布になっていると考えられる。

次に、信頼性係数と測定の標準誤差を示す。リスニングセクションの信頼性係数は、K-R 21が.01、K-R 20が.16であった。それぞれの信頼性係数を用いた測定の標準誤差は、2.23と2.06であった。

次に、項目容易度と項目弁別力について述べる。表２は、ミシガンテストのリスニングセクションの各テスト項目の項目容易度と項目弁別力を示している。項目容易度は、.13（#20）から.80

表１．基礎統計量

項　目	統計値
平均	9.39
標準偏差	2.24
95% 信頼区間	
下限	8.94
上限	9.84
最小値	4.00
最大値	16.00
歪度	0.29
標準誤差（歪度）	0.24
尖度	0.29
標準誤差（尖度）	0.48

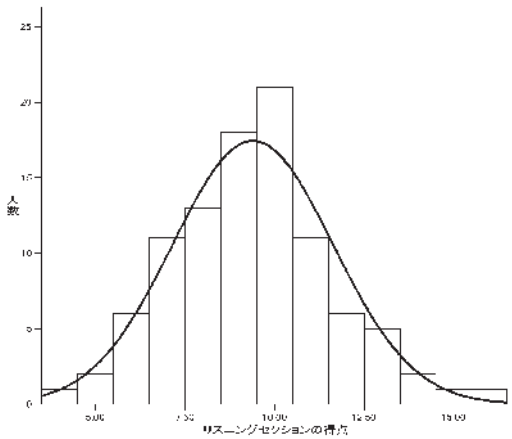

図１．リスニング得点分布

表２．ミシガンテスト・リスニングセクションの項目容易度と項目弁別力

項目#	項目容易度	項目弁別力	項目#	項目容易度	項目弁別力
#01	0.60	0.33	#11	0.37	0.21
#02	0.71	0.18	#12	0.18	0.21
#03	0.15	0.03	#13	0.52	0.18
#04	0.62	0.42	#14	0.72	0.09
#05	0.52	0.27	#15	0.36	0.21
#06	0.34	0.18	#16	0.42	0.30
#07	0.48	0.42	#17	0.31	0.36
#08	0.41	0.24	#18	0.74	0.33
#09	0.59	0.42	#19	0.80	0.00
#10	0.41	0.33	#20	0.13	-0.03

（#19）の範囲になった。項目容易度が.90を超えるような極端に簡単すぎる項目や、項目容易度が.10以下である極端に難しすぎる項目はみられなかった。また、適切な範囲（.30～.70）の項目は、20項目中13項目あった。

項目弁別力については、次のような結果となった。.40以上のよい項目は、３項目であった（#04、#07、#09）。.30から.39の改良が必要かもしれないよい綱目は、５項目であった（#01、#10、#16、#17、#18）。.20から.29の改良が必要な項目は、５項目であった（#05、#08、#11、#12、#15）。.19以下の項目弁別力のよくない項目は、７項目であった（#02、#03、#06、#13、#14、#19、#20）。

項目容易度が適切な範囲（.30～.70）で、かつ項目弁別力が.30以上の項目は、７項目であった（#01、#04、#07、#09、#10、#16、#17）。項目弁別力の下限を.20以上に設定すると、４項目（#05、#08、#11、#15）増え、11項目が含まれる。

３．２．ラッシュ分析

ラッシュ分析から得られる困難度推定値をみると、最も難しいテスト項目（#20）の困難度推定値は1.83ロジッツであり、最も簡単なテスト項目（#19）の困難度推定値は-1.61ロジッツであった（表３参照）。すなわち、ミシガンテストのリスニングセクション20項目の分布の範囲は、3.44ロジッツ（-1.61～1.83）であり、十分な散布度を示していることがわかる（図２参照）。

また、最も難しい２つの項目のロジッツ値はそれぞれ1.83と1.65であり、最も高い能力の参加者のロジッツ値は1.63であった（表４参照）。これは、最も能力の高い学習者であっても、正答する可能性が50％以下である項目が存在していたことを示している。すなわち、項目の困難度の方が、学習者の能力よりも高く、天井効果を避けることができるといえる。このことは、項目の困難度推

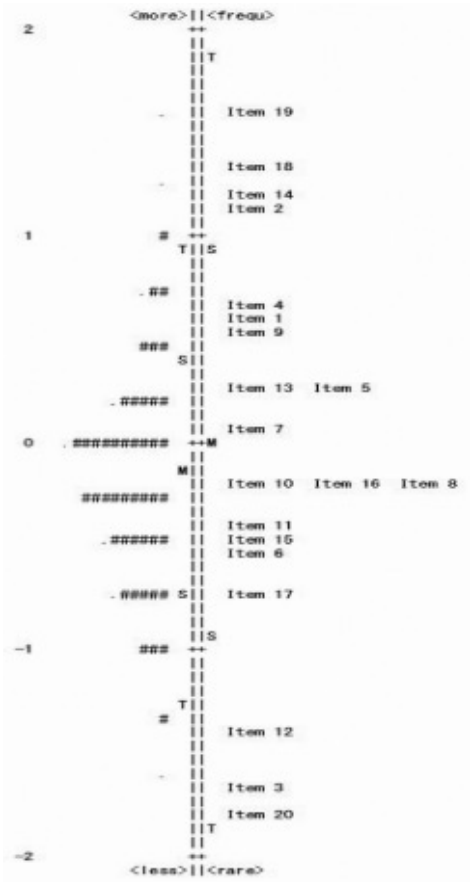

図２．Item-Person Map

定値の範囲が妥当であることを示す。一方、最も容易な項目のロジッツ値は-1.61であったのに対して、それより低い能力（-1.63ロジッツ）の学習者が１名いた。参加者数が98人であったことを考えると、学習者の能力推定値の範囲とテスト項目の困難度推定値の範囲はほぼ同じであったといえよう。

しかしながら、困難度の推定値の分布に問題も指摘できる。それは、項目同士が離れていることである（図２参照）。最も大きな推定値の差は、#17と#12の間の.71であった。２番目に大きな差は、＃４と＃２の間の.44であった。これらの項目間の空白部分に、多くの参加者たちのリスニング能力の推定値が位置している。このことは、学習者の能力の測定に誤差をもたらしてしまっている。今後改良するとすれば、困難度推定値に間隔のみられるレベル（.00、.05、.10、.30、.60、.80～1.30、-.30、-.40、-.80～-1.10）の困難度の推定値を持つ項目を追加することである。このレベルの項目は、参加者のリスニング能力に関する有益な情報を提供すると同時に、測定の誤差を減少させることになると思われる。

次に、テスト項目の誤差の推定値をみる。Wright（1977）は、サンプルサイズが100名程度のと

表３．項目困難度推定値得点分布

					インフィット		アウトフィット	
項目 #	素点	*n*	推定値	標準誤差	平方平均	*t* 統計値	平方平均	*t* 統計値
#20	13	98	1.83	0.30	1.01	0.10	1.30	1.10
#03	15	98	1.65	0.29	1.11	0.60	1.23	1.00
#12	18	98	1.42	0.27	0.98	-0.10	0.95	-0.20
#17	30	98	0.71	0.23	0.96	-0.40	0.92	-0.70
#06	33	98	0.56	0.22	1.00	0.00	0.99	-0.10
#15	35	98	0.46	0.22	1.02	0.20	1.01	0.10
#11	36	98	0.42	0.22	0.99	0.00	0.98	-0.20
#08	40	98	0.23	0.21	1.01	0.10	1.01	0.10
#10	40	98	0.23	0.21	0.94	-1.00	0.94	-0.90
#16	41	98	0.19	0.21	1.02	0.40	1.04	0.60
#07	47	98	-0.08	0.21	0.92	-1.70	0.90	-1.70
#05	51	98	-0.25	0.21	1.00	0.00	1.00	0.00
#13	51	98	-0.25	0.21	1.01	0.30	1.02	0.30
#09	58	98	-0.56	0.21	0.95	-0.70	0.94	-0.90
#01	59	98	-0.61	0.21	0.95	-0.80	0.93	-1.00
#04	61	98	-0.70	0.22	0.99	-0.20	0.96	-0.40
#02	70	98	-1.14	0.23	1.05	0.50	1.09	0.70
#14	71	98	-1.20	0.23	1.08	0.80	1.19	1.40
#18	73	98	-1.31	0.24	0.94	-0.50	0.88	-0.80
#19	78	98	-1.61	0.26	1.06	0.40	1.25	1.30
M	46.0		0.00	0.23	1.00	-0.10	1.02	0.00
SD	18.9		0.94	0.03	0.05	0.60	0.12	0.80

き、誤差の推定値の基準として.25ロジッツを挙げている。本研究では、20項目中16項目のテスト項目の誤差の推定値が.21ロジッツから.24ロジッツの間に含まれていた。Wright（1977）の基準に従えば、リスニングセクションのほとんどの項目において、誤差の推定値が小さかったといえる。

リスニングセクションにおける20項目の適合統計量（fit statistics）は、すべて許容できる範囲内（0.8～1.2）におさまっていた。これは、ラッシュモデルが予測する値との適合度が高いことを示している。リスニングセクション全体において、ミスフィット（misfit）はみられなかった。モデル適合度がよかったことは、テスト項目が有効に機能していたことを示していると同時に、ミシガンテストのリスニングセクションの一元性（unidimensionality）を示唆するものである。

表４．受験者能力推定値（上位・下位９名ずつ）

					インフィット		アウトフィット	
ID #	素点	*k*	推定値	標準誤差	平方平均	*t* 統計値	平方平均	*t* 統計値
70	16	20	1.63	0.60	1.18	0.60	1.92	1.50
68	15	20	1.30	0.56	1.39	1.30	1.92	1.80
29	14	20	1.00	0.53	0.86	-0.50	1.01	0.10
78	14	20	1.00	0.53	0.95	-0.10	0.82	-0.40
28	13	20	0.73	0.51	0.84	-0.70	0.74	-0.90
41	13	20	0.73	0.51	0.86	-0.60	0.85	-0.40
50	13	20	0.73	0.51	0.65	-1.80	0.55	-1.70
65	13	20	0.73	0.51	0.69	-1.50	0.61	-1.40
90	13	20	0.73	0.51	0.90	-0.40	0.95	-0.10
9	6	20	-1.01	0.53	0.64	-1.70	0.53	-1.40
36	6	20	-1.01	0.53	0.94	-0.20	0.90	-0.20
44	6	20	-1.01	0.53	0.94	-0.20	0.92	-0.10
55	6	20	-1.01	0.53	1.25	1.10	1.71	1.70
66	6	20	-1.01	0.53	1.43	1.70	1.92	2.10
92	6	20	-1.01	0.53	0.95	-0.10	1.47	1.20
59	5	20	-1.30	0.55	0.79	-0.70	1.36	0.90
94	5	20	-1.30	0.55	1.07	0.30	1.03	0.20
53	4	20	-1.63	0.59	1.02	0.20	0.76	-0.30
M	9.4		-0.16	0.50	0.99	0.00	1.02	0.10
SD	2.2		0.56	0.02	0.20	1.00	0.31	1.00

４．考察

本研究において、ミシガンテストのリスニングセクションが、日本の大学レベルにおける英語学習者の熟達度を測定するための適切性について検証した。本節では、２つのテスト分析方法の結果について考察する。

古典的項目分析においては、ミシガンテストのリスニングセクションが本研究の参加者の熟達度を測定するのに適していることがわかった。平均値（9.39）は、得点の可能範囲（０～20）のほぼ中央に位置していた。また、標準偏差（2.24）をみると、得点の分布にわずかな制約があったことがわかる。しかし、この点は、歪度及び尖度の点から正規性の逸脱が本データには認められないことが示唆されるため、特に問題ではないと思われる。さらに、古典的項目分析の結果によれば、極

端に困難度が高い項目も、極端に困難度が低い項目もなかったことがわかった。つまり、当該テストの項目は参加者にとって適切な難易度であったと言えよう。

しかしながら、項目容易度と項目弁別力の点から許容範囲内に入っていた項目は、20項目中７項目（35%）だけであった。もし項目弁別力の下限を.20に設定すると、20項目中11項目（55%）が好ましい項目であると見なされることになる。先行研究をみてみると、好ましい項目とみなされる項目について本研究と同様の割合を示している。例えば、Brown（1989）は、60項目中35項目（58%）が好ましい基準を満たすとした。Culligan＆Gorsuch（1999）は、150項目中66項目（44%）が項目弁別力の下限である.20を満たしているとした。また、Westrick（2005）は、120項目中46項目（38%）が、好ましい項目容易度の基準（.30～.70）を満たしたと報告した。さらに、Sakai＆Wistner（in press）は、学内用のプレイスメントテストを分析し、30項目中15項目（50%）が好ましい項目であると報告した。これらの先行研究と比較すると、本研究における好ましい項目の割合が特に低かったとはいえないと考えられる。

古典的項目分析において、最も問題があると考えられる結果は信頼性係数であった。本研究では、K-R 21が.01、K-R 20が.16であった。古典的なテスト理論においては、信頼性の推定値は、真の信頼性よりも低く推定される傾向があると考えられている。しかし、この点を考慮に入れても、本研究の信頼性係数の低さは問題であろう。信頼性係数が低かったことの理由として、標準偏差が小さかったことが考えられる。高い信頼性係数の推定のためにはある程度の分散が必要である。本研究のデータは正規分布を逸脱していないが、十分な信頼性の値を得るほどには得点が散布していなかったと言えよう。テストの得点が十分に分布していないことに加えて、テスト項目数が少なかったこと（20項目）が要因として考えられよう。信頼性はテスト項目数の影響も受ける。そのため、十分な信頼性係数を得るほど、リスニングセクションにおけるテスト項目の数が多くなかったと言える。

この点は、検討が必要だと思われる。ミシガンテストには、リスニングセクションの他に、文法、語彙及び読解セクションが含まれる。しかし、項目数を増やすために、別の能力を測定している他のセクションを含めて、ミシガンテスト全体として信頼性係数を計算することには問題があると考えられる。むしろ、テストの信頼性係数の低さを解決する方法として、リスニングセクションのテスト項目の種類を増やすことが挙げられる。ミシガンテストには、ＡからＣまでの３つのフォームがあるが、少なくともリスニングテストに関しては、３つのフォームを同時に与えることも１つの解決方法であろう。

ラッシュ分析においても、ミシガンテストのリスニングセクションは参加者の熟達度を測定するために適度な難易度であるという結果が出た。まず、困難度推定値のばらつきの点から項目同士に間隔があるという問題点はあったが、テスト項目の困難度推定値は参加者の能力推定値よりも広い範囲にわたって分布しており、また両者は分布範囲も類似していた。次に、項目の誤差の推定値が低かったことがわかった。20項目中16項目が.25以下の誤差の推定値であり、残りの４項目も.26～.30以内の数値であった。全ての項目の誤差の推定値は妥当な範囲であったといえよう。３つ目

として、適合度の点からリスニングセクションのテスト項目が有効に機能していると言える。20項目全てに関して、ラッシュ分析が予測するモデルと合致することがわかった。

５．おわりに

本研究は、ミシガンテストのリスニングセクションが、日本の大学レベルにおける英語学習者の熟達度を測定するのに適切であるかどうかに関して、テスト項目の機能と難易度の点から統計的に分析した。テスト分析方法として、古典的項目分析法とラッシュ分析法が用いられた。どちらの分析方法からも本研究の参加者の熟達度を測定するのに、有益な情報を得ることができた。

今後は、別の日本人英語学習者のグループを対象とした研究が必要であると考えられる。本研究の参加者は、上級クラスの学生に限れていた。もし、別のクラスの学生を含めたときに、古典的項目分析やラッシュ分析法によって項目の難易度と参加者の能力が合致するのかどうか、信頼性係数を高めることができるのか、など検討する必要があろう。また、本研究では、熟達度テストの使用目的については、限定しなかった。例えば、プレイスメントを目的とするミシガンテストの使用は適切なのかどうかは、今後の課題であろう。

（あべ　まりこ・本学経済学部准教授／ブライアン　ウィスナー・東京純心女子大学現代英語学科講師／
さかい　ひでき・信州大学教育学部准教授）

引用文献

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

Brown, J. D. (1989). Improving ESL placement tests using two perspectives. ***TESOL Quarterly, 23,*** 65-83.

Brown, J. D. (2005). ***Testing in language programs: A comprehensive guide to English language assessment*** (New ed.). New York: McGraw Hill.

Corrigan, A., Dobson, B., Kellman, E., Spaan, M.,＆Tyma, S. (1993). ***English placement test.*** The Testing and Certification Division, English Language Institute, University of Michigan.

Culligan, B.,＆Gorsuch, G. (1999). Using a commercially produced proficiency test in a one-year core EFL curriculum in Japan for placement purposes. ***JALT Journal, 21,*** 7-28.

Field, A. (2005). ***Discovering statistics using SPSS*** (2nd ed.). London: Sage publications.

Fulcher, G. (1997). An English language placement test: Issues in reliability and validity. ***Language Testing, 14,*** 113-138.

Gorsuch, G. J.,＆Culligan, B. (2000). Using item response theory to refine placement decisions. ***JALT Journal, 22,*** 315-325.

JLTA Language Testing 用語集編集委員会．(2006)．『日本言語テスト学会　言語テスティング用語集―日本言語テスト学会　テスティングの実施規範』長野：日本テスト学会．

Sakai, H.,＆Wistner, B. (in press). Classical item analysis of an in-house English placement test: Issues in appropriate item difficulty and placement precision. ***JACET Chubu Journal, 5.***

Westrick, P. (2005). Score reliability and placement testing. ***JALT Journal, 27,*** 71-93.

Wright, B. (1977). Solving measurement problems with the Rasch model. ***Journal of Educational Measurement, 14,*** 97-116.

靜哲人．(2007)．『基礎から深く理解するラッシュモデリング―項目応答理論とは似て非なる測定のパラダイム』大阪：関西大学出版部．