— Article —

ラッシュモデルによるTOEIC®の構成概念的妥当性検証
Rasch Model-Based Construct Validation for TOEIC® Test

伊藤彰浩（西南学院大学）
Akihiro ITO
Seinan Gakuin University
e-mail: ito@seinan-gu.ac.jp

Abstract

The present study investigated the possibility of a Rasch model-based construct validation of standardized English proficiency tests from the perspectives of the six facets of Messick's (1989) *unitary* construct validity framework. The six facets are (1) content, (2) substantive, (3) structural, (4) generalizability, (5) external, and (6) consequential. This paper provides a brief historical overview on the definition of validity in psychometrics and then states the application of the Rasch model for language test validation. We highlight the primary advantage of the Rasch model: yielding a hypothetical unidimensional line along which items and persons are located according to the item difficulties and ability measures. Recent research findings in psychometrics have implied that the principles of the Rasch model can be related to the Messick's (1989) construct validity issue. To examine the potentiality of linking the Rasch model with Messick's (1989) validity framework in language testing research context, TOEIC® data for 136 students learning English as a foreign language in Japanese universities were gathered and statistically analysed by the *WINSTEP Rasch Measurement Computer Program* (Linacre, 2008). The following are the results: (1) the TOEIC® test showed satisfactory separation values and item strata; (2) there were several misfit items to the model; (3) a few items had distractors that did not function in the way they had been intended to; (4) the item-person map showed that the TOEIC® test was on-target and covered a wide range of ability scale; and (5) the issue of unidimensionality remained undetermined because of the mixed and hard-to-interpret results from Rasch measures and the principle component analysis on the residuals (PCAR). The paper concludes that the Rasch model can be a powerful tool for establishing the validity of language tests. Directions for future discourse based on the *Messick–Rasch Link* are also discussed.

Keywords: language testing, validity, facet, Rasch model, Messick–Rasch Link

緒言

本論文の目的はラッシュモデルを利用して TOEIC®の構成概念的妥当性 (construct validity) を検証することである。本論文では初めに妥当性 (validity) の定義に関する歴史的変遷をたどりながら、Messick (1989) の妥当性理論の成立過程を説明する。次に構成概念的妥当性検証の方法としてのラッシュモデルの利用価値について検討す

る。そして、日本の大学で学ぶ136名の英語学習者から得られたTOEIC®のデータをラッシュモデルにより分析する。分析結果を Messick (1989) が提唱した妥当性の 6 つの側面（(1) 内容的側面 (content facet)、(2) 本質的側面 (substantive facet)、(3) 構造的側面 (structural facet)、(4) 一般化可能性 (generalizability)、(5) 外的側面 (external facet)、(6) 結果的側面 (consequential facet) ）の観点から考察し、TOEIC®の構成概念的妥当性の全体像に迫る。

妥当性の定義に関する歴史的変遷

テストの妥当性の概念は時代によって変化してきた。妥当性の概念に初めて言及した Kelly (1927) は、測定したいものが測定できるテストは妥当なテストだと論じた。この定義が広く認められた時期は、操作主義 (manipulationism) の時代として位置づけられている。この時代には測定の方法 (method) が測定対象の概念を規定することが暗黙の了解とされ、尺度 (scale) が何を測定しているかは検討の対象から外された。その結果、外部基準 (external criterion) として認められる指標との相関関係の程度によって、テストの妥当性は検証された(Anastasi, 1954)。

1954 年にはアメリカ心理学会 (American Psychological Association) がテストの妥当性を(1)内容的妥当性 (content validity)、予測的妥当性 (predictive validity)、併存的妥当性 (concurrent validity)、構成概念的妥当性 (construct validity) の 4 つに分類した (American Psychological Association, 1954)。Cronbach & Meehl (1955) もこの分類にしたがって研究を行う意義について論じているが、その中でも特に構成概念的妥当性の重要性を主張し、論理実証主義 (logical positivism) に基づく法則定立ネットワーク (nomological network) の理論的枠組みを提唱した。この理論的枠組みでは、尺度は理論的または仮説的な構成概念を測定するものと想定され、テストの構成概念的妥当性の検証はテスト・データを利用した構成概念間の関係を追究する一連の活動であるとされた。同時期には、Cambell & Fiske (1959) によって収束的妥当性 (convergent validity) と弁別的妥当性 (discriminant validity) の概念が提唱され、構成概念的妥当性の検証法として多特性･多方法相関行列法 (multitrait-multimethod correlation matrix) が提案された。この時代の構成概念的妥当性の検証は、収束的妥当性 (convergent validity) と弁別的妥当性 (discriminant validity) の概念を利用し、様々な尺度間の相関係数を測定し、その結果に基づいた仮説検証の活動であったと言える。

1960 年代に入るとアメリカ心理学会がテスト基準を変更し、併存的妥当性 (concurrent validity) と予測的妥当性 (predictive validity) を合わせた概念として基準関連的妥当性 (criterion-related validity) が提唱された (American Psychological Association, 1966)。その結果、テストの妥当性は、基準関連妥当性、内容的妥当性、構成概念的妥当性の 3 つに分類されることになった。この分類は広く浸透したが、3 つの妥当性の概念が別々に検証される結果となり、それぞれの妥当性の検証結果を統合し、テストの妥当性の全体像に迫る研究活動は衰退することになった (Landy, 1986)。その反省として、3 つの妥当性の概念を統合しようとする活動が盛んになり、Messick (1989) は妥当性の概念の単一性 (unitary nature) を主張し、構成概念的妥当性こそが

テストの本質的な妥当性であると論じた。

Messick (1989) は構成概念的妥当性を単一的な概念とする一方、そこには 6 つの側面 (facet) があると論じている。それらは (1) 内容的側面 (content facet)、(2) 本質的側面 (substantive facet)、(3) 構造的側面 (structural facet)、(4) 一般化可能性 (generalizability)、(5) 外的側面 (external facet)、(6) 結果的側面 (consequential facet) である。以下、各側面の内容を列挙する。

(1) 内容的側面：テストを構成する各テスト項目は測定対象である能力を反映しているか。

(2) 本質的側面：内容的側面が実際のテストの解答時に受験者の具体的な活動として具現化されているか。

(3) 構造的側面：テストは実施された時点においてひとつの能力を測定しているか。すなわち、テストの一次元性 (unidimensionality) が確認できるか。

(4) 一般化可能性：実施されたテストの得点とその解釈によって得られた判断が、テストが実施された時点や場所以外の状況でも一般化できるか。

(5) 外的側面：受験者の知識や能力が変化した場合、その変化を得点として表す性質をテストが備えているか。

(6) 結果的側面：テストの得点が正当に評価され、社会の中で行われる意思決定や価値判断に対して利用された場合、多くの人が納得できる結果をもたらすことができるか。

上記の Messick (1989) の妥当性の概念は、テストに関係する心理測定学を初めとし、言語テスト研究 (language testing) においても広く浸透している (McNamara, 1996)。[1] この枠組みに基づいてテストの妥当性を検証しようとする場合、6 つの側面それぞれに対して焦点を当てるため複数の研究法を採用しなければならない。なぜなら一般化可能性と結果的側面の例から分かるように、Messick (1989) の妥当性の概念はテスト・データ自体の属性に留まらず、テストが利用される具体的な場 (context) において検証されるべき課題も含まれているからである。しかし、複数の研究法を採用し、妥当性のそれぞれの側面を別々に検証するのであれば、Messick (1989) が妥当性の単一性を主張する以前に行われた複数の妥当性を別々に検証するアプローチと何ら変わりがない。妥当性の単一性の概念に基づいてテストの妥当性を検証するのであれば、妥当性の全体像を把握でき、同時に 6 つの側面に対する判断や解釈を可能にする単一の方法が選定されるべきである。すなわち妥当性の概念とその検証法の間に論理的対応関係が成立しなくてはならない。換言すれば、テスト・データを Messick (1989) の妥当性の概念に「つなげる」(linking) 単一の理論的枠組みが必要とされる (Bond, 2003)。その理論的枠組みとして近年、脚光を浴びているのが統計モデルのひとつであるラッシュモデル (Rasch model)である。Bond (2003) はラッシュモデルを利用した Messick (1989) の妥当性の概念に基づくテストの構成概念的妥当性の検証を「ラッシュモデルと Messick の妥当性概念の連結」(Rasch-Messick Link) と呼称し、Messick (1989) によって提唱された妥当性の「理論」(theory) と心理測定学による「実践」(practice) の融合であると主張している (pp. 179-181)。

ではラッシュモデルとはどのような統計モデルなのだろうか。そしてラッシュモデルをどのように利用すれば Messick (1989) の妥当性の概念に基づくテストの妥当性検証を行うことができるのだろうか。次節ではラッシュモデルとラッシュモデルを利用したテストの構成概念的妥当性検証について検討する。

ラッシュモデルとテストの構成概念的妥当性検証

ラッシュモデルはテスト受験者の解答が、ガットマンの応答パターン (Guttman response pattern) (Guttman, 1950) に対応することを前提とした規範的確率統計モデルである (Rasch, 1960)。このモデルでは、受験者の能力が高くなるほど、テストにおける正答数が徐々に増加し、反対に、学習者の能力が低くなるほど、正答数が徐々に減少することが想定されている。素点 (raw score) に基づく古典的テスト理論 (classical test theory) とは異なり、ラッシュモデルは、自然対数によって換算するロジェット得点 (logit score) により絶対的な原点を算出し、受験者のテスト項目への応答パターンを予測する。そして、ロジェット得点からテスト項目の難しさと受験者の能力の程度を算出することもできる。絶対的な原点を持つ受験者測定値 (person measure)と項目測定値 (item measure) によって、同一の尺度に沿って受験者の能力を推定でき、受験者集団に依存しない項目特性の算出が可能となる（島谷・法月・伊藤・木下、2009a; Shimatani, Norizuki, Ito & Kinoshita, 2009b)。

では、ラッシュモデルを利用して Messick (1989) の妥当性の概念における 6 つの側面をどのように検討できるのだろうか。Bond & Fox (2001)、Smith, Linacre & Smith (2003)、Baghaei (2008, 2009, 2010) はラッシュモデルを利用した際に得られる各種の測定値が Messick (1989) の構成概念的妥当性の 6 つの側面に結び付くと論じている。Baghaei (2010) が利用したラッシュ分析ソフトウェア *WINSTEP* (Linacre, 2008) によって得られる測定値の名称を挙げながら解説する。

WINSTEP (Linacre, 2008) によって得られる測定値は、 (1) 分離指数 (separation indices)、(2) 項目階層値 (item strata) (3) 項目フィット係数 (item fit indices)、(4) 受験者フィット係数 (person fit indices)、(5) 受験者・項目散布図 (person-item map)、(6) 項目測定値 (item measure)、(7) 受験者測定値 (person measure)、(8) 項目・測定値相関係数 (item-measure correlation)、そして (9) 錯乱肢統計値 (distractor statistics)の 9 種類である。[2] 以下に各側面と測定値の関係を示し、解説を加える。

(1) 内容的側面：項目フィット係数、受験者・項目散布図、項目階層値
　　項目・推定値相関係数
(2) 本質的側面：受験者フィット係数、錯乱肢分析
(3) 構造的側面：項目フィット係数、受験者フィット係数
(4) 一般化可能性：項目推定値、受験者推定値
(5) 外的側面：受験者・項目散布図
(6) 結果的側面：受験者・項目散布図

内容的側面は、テストを構成する各テスト項目が測定対象である能力を反映しているかに関連している。この問題に関する情報は項目フィット係数、受験者・項目散布図、項目階層値、項目・推定値相関係数から得ることができると考えられる (Baghaei, 2008)。なぜなら項目フィット係数はテスト項目と測定を意図された構成概念との関連性の程度を示すからである。ラッシュモデルに適合しないミスフィット項目は測定を意図された構成概念以外を測定対象としていると考えられるため、その様な項目が多い場合は、内容的側面に問題があり当該項目の削除、修正が必要となる。さらに受験者・項目散布図から、テストにおいて万遍なく難度レベルの異なるテスト項目が配置され、意図された測定対象の内容がテスト全体によってカバーされているかを知ることができる。項目・測定値相関係数は特定のテスト項目における得点が他の項目の得点とどの程度の関係があるかを示していることから、相関係数は正の値を示すべきである。そして、項目階層値はテスト項目の難度階層の数を具体的に示すため、受験者・項目散布図から検討されるテスト項目の配置状況を確認するために有効な指標である。

本質的側面は、内容的側面が実際のテストの解答時に受験者の具体的な活動として具現化されているかに関連している。これに関する情報は受験者フィット指数によって得ることができると考えられる (Baghaei, 2010)。受験者フィット指数は受験者個人のテスト項目への応答パターンがモデルの予測とどの程度適合しているかを示している。受験者フィット係数の値が低い場合、受験者の疲労、不注意や推測に依存した解答パターンが考えられる。さらに研究対象とするテストが多肢選択方式の場合は、錯乱肢分析を行うことによって、能力の高い受験者ほど錯乱肢に惑わされずに正解を選択し、能力が低くなるにつれて錯乱肢を選択する傾向があるか調査することも可能である (Baghaei, 2010)。

構造的側面は、テストの一次元性に関連している。テスト・データの一次元性に関する情報は、受験者フィット係数により受験者の全般的な解答パターンがモデルに適合しているか、そして項目フィット係数により各々の項目が同一の能力を測定しているかの両面から検討できる (Bond & Fox, 2001)。さらにラッシュモデルで得られた測定値によるデータの分散説明率を求め、その値が低いと判断される場合は残差主成分分析を行い、その結果に基づいて一次元性に関する最終的な判断を行うことも可能である (Goh & Aryadoust, 2010)。

一般化可能性はテスト項目の不変の程度、すなわちテストの一次元性や信頼性の概念に関係している。したがって、項目推定値と受験者推定値を利用し、推定値全般の安定性を検討することで、一般化可能性の程度について検討を加えることができる (Baghaei, 2009)。

外的側面は、同一のテストを同じ学力レベルの学習者集団に実施し、一方の学習者集団に何らかの形式教授 (formal instruction) を行った後に、テストを再度実施した場合、その学習効果がテスト得点として反映されるか調査することで検証が可能となる。しかし、ラッシュモデルを利用した場合は、テストを2度実施する研究デザインを前提としていないので、受験者・項目散布図を利用することで、受験者能力測定値に対

して適切な難度の項目が配置され、受験者の能力が変化した場合でも、その能力の変化を推定できる範囲に項目が配置されているかによって検討できる (Bond & Fox, 2001; Bahaei, 2010)。

結果的側面の検討に直接的につながる情報をラッシュモデルが提供しているとは考えにくいかもしれない。しかし、テスト得点に基づく判断や意思決定が正確に行われるためには、テスト項目が各種のバイアスや不自然な機能による影響をどの程度受ける可能性があるかについて総合的な判断を行えば、テストが正しい意思決定や評価に結び付く必要条件を備えているか検討できる (Baghaei, 2010)。したがって、外的側面と同様に受験者・項目散布図の結果から判断できると考えられる (Bond & Fox, 2001)。

これまでの説明から分かるように、規範的確率統計モデルのひとつであるラッシュモデルで得られる測定値には妥当性の側面に関連する情報が含まれていると考えられる。次節では、ラッシュモデルを利用して英語標準テストのひとつである TOEIC® の構成概念的妥当性を検証する意義について論じる。

ラッシュモデルを利用した TOEIC®の構成概念的妥当性検証の意義

近年、ラッシュモデルを利用し、テスト・データの分析を行う研究は国内外において数多く行われている。我が国における日本人英語学習者を対象とした言語テスト研究においてもラッシュモデルの利用は拡大を続けている（例：大友、1996; 大友・中村、2002; 靜、2007; 阿部・ウィスナー・酒井、2008; 島谷・法月・伊藤・木下、2009a）。しかし、Messick (1989) の妥当性に関する理論的枠組みとラッシュモデルに基づく分析結果を連結し、その解釈によってテストの構成概念的妥当性にアプローチする研究の数は限定的であり、筆者の知る限り日本の英語教育において利用されている標準テストを対象とした研究は実施されていない。そこで本研究では日本社会の様々な場面で用いられている英語テストである TOEIC®を研究対象とする。TOEIC®は現在の日本では個人による受験のほか 2,500 以上の企業、団体、学校などで昇任条件、単位認定、入学試験、クラス編成のために利用されている high-stakes な英語学力テストである（伊藤・島谷・法月・木下、2009a）。その一方で実際の英語教育の場における TOEIC®の妥当性に関する研究は未だ充分に行われておらず、現在、様々な観点からの研究が継続的に実施されている（例：伊藤・島谷・法月・木下、2009b, 2010; 島谷・法月・伊藤・木下、2009a; Norizuki, Ito & Shimatani, 2011）。Messick (1989) の妥当性の 6 つの側面から TOEIC®の妥当性を検証することで、これまでの研究では明らかとなっていない TOEIC®の特性に関する総合的な情報を得ることが期待できる。

以上の議論から、本研究では日本の大学で学ぶ学生を対象に実施された、英語標準テストである TOEIC®のデータを利用し、構成概念的妥当性の検証を行う。調査課題は Messick (1989) が示した妥当性の 6 つの側面の内容に基づいて以下のように設定する。

研究課題

(1) 内容的側面：TOEIC®の各テスト項目は測定対象とする能力を反映しているか。
(2) 本質的側面：TOEIC®の各テスト項目への応答パターンはラッシュモデルの予測に適合しているか。
(3) 構造的側面：TOEIC®は実施された時点でひとつの能力を測定しているか。すなわち、テストの一次元性が確認できるか。
(4) 一般化可能性：TOEIC®のテストの得点によって行われた判断は、テストが実施された時点や場所以外の状況でも一般化できる可能性を示しているか。
(5) 外的側面：TOEIC®は、受験者の英語能力が変化した場合、その変化を得点として表す性質を有しているか。
(6) 結果的側面：TOEIC®は、社会の中で行われる意思決定や価値判断に対して利用された場合、多くの人が納得できる結果をもたらす可能性があるか。

方法

調査目的

本論文の目的は Messick (1989) が提唱したテストの構成概念的妥当性の 6 つの側面に着目し、ラッシュモデルに基づいて TOEIC®の構成概念的妥当性の検証を行うことである。

調査参加者

本調査の参加者は英語を外国語として学び日本の大学に在籍する 136 名（日本人学生 115 名、留学生 21 名（中国 8 名、バングラディシュ 4 名、インドネシア 3 名、ネパール 3 名、ミャンマー1 名、ホンジュラス 1 名、フィリピン 1 名））であった。調査参加者が在籍する大学は、西南学院大学、熊本大学、静岡産業大学、長崎純心大学、東京情報大学であった。調査参加者の所属学部は、文学部、法学部、商学部、経済学部、国際文化学部、教育学部、情報学部、理学部、工学部、薬学部、医学部、人文学部、人間科学部、総合情報学部、大学院教育学研究科であった。

調査材料

TOEIC® 運営委員会が出版している TOEIC®公式問題集の中から、新 TOEIC®問題集 (TOEIC®運営委員会、2007) に含まれる「テスト II」を使用して、TOEIC® Practice Test のテスト冊子を作成した。テスト形式は一貫して多肢選択方式であり、聴解セクションと読解セクションそれぞれ共に 100 問の計 200 問で構成されている。公式問題集の TOEIC® Practice Test は、実際の TOEIC®と同様の質と量であることが国際ビジネスコミュニケーション協会によって認定されている。したがって、本研究で得られるテスト・データは実際の TOEIC®から得られるデータと同一であると想定される。本研究で利用した TOEIC®のセクション、パート、パート内容、セクション、問題数（項目数）および解答時間は表 1 に示されている通りである。

調査手順

TOEIC®は本論文執筆者および研究協力者の所属校で実施された。テスト実施日は2008年11月から12月であった。調査参加者となった136名は、TOEIC® Practice Testを自ら申し込んだ。テストは2008年11月から12月に実施された。テストの実施時間は実際のTOEIC®と同じ聴解セクション45分、読解セクション75分の計2時間であり、最初の説明およびテスト制限時間と終了後のアンケートの時間を含めて約3時間であった。受験者はマークシート（教育ソフトウェア社 総合カード103）に回答を記入した。マークシートはテスト終了後に回収され、マークリーダー（教育ソフトウェア社 SR-430）によってデータの読み取りとコンピュータへの入力が行われた。ラッシュモデルによる分析には *WINSTEP Rasch Measurement Computer Program (Version 3.66.0)* (Linacre, 2008) が利用された。

表1 TOEIC®の問題構成

TOEIC®		
パート	パート内容	問題数
聴解セクション（45分間）		
1	写真描写問題	10
2	応答問題	30
3	会話問題	30
4	説明文問題	30
読解セクション（75分間）		
5	短文穴埋め問題	40
6	長文穴埋め問題	12
7	読解問題 Single passage Double passage	 28 20

(Mahoney, Matsuura & Murakami, 2007)

結果と考察

ここでは最初にラッシュモデルによる分析から得られた測定値を提示する。[3] そして構成概念的妥当性の6つの側面それぞれに関連する測定値と分析結果を取り上げながら各側面について考察を加える。

表2はTOEIC®受験者136名の基本測定値、表3はTOEIC®のテスト項目（200項目）の基本測定値を示している。Linacre (2008) による *WINSTEP* のマニュアルによれば、受験者分離係数 (PERSON SEPARATION) は 2.00 以上、受験者信頼性係数 (PERSON RELIABILITY) は.80以上、項目分離係数 (ITEM SEPARATION) は3.00 以上、項目信頼性係数 (ITEM RELIABILITY) は.90以上[4]あれば、研究対象としたテストは適切に機能したと判断できると考えられている。受験者分離係数はテストが高い能力と低い能力の受験者を識別できる度合いを示し、項目分離係数は受験者集団が項目難易度の階層を識別できる度合を示していると解釈できる。

表2と表3の結果から、受験者分離係数は3.85、受験者信頼性係数は.94、項目分離係数は5.13、項目信頼性係数は.96を示したことから、すべての値が基準を上回った。さらにこれらの値はラッシュモデルが予測する数値 (MODEL) とも近接していることからも、4つの指標は適切な値を示していると言える。

表4はTOEIC®の項目フィット係数を示している。実際の分析データではTOEIC®の200項目すべてのデータが提示されるが、紙数上の都合でインフィット項目 (INFIT) とアウトフィット項目 (OUTFIT) と判断された項目の情報のみ掲載した。

表2　基本測定値（TOEIC®受験者136名）

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	119.0	200.0	.00	.17	1.00	.0	.99	.1
S.D.	23.3	.0	.69	.02	.09	1.3	.17	1.1
MAX.	191.0	200.0	3.09	.35	1.27	3.6	1.57	3.9
MIN.	71.0	200.0	-1.31	.16	.82	-2.7	.51	-1.8
REAL RMSE	.17	TRUE SD	.67	SEPARATION	3.85	PERSON RELIABILITY		.94
MODEL RMSE	.17	TRUE SD	.67	SEPARATION	3.92	PERSON RELIABILITY		.94
S.E. OF PERSON MEAN = .06								

表3　基本測定値（TOEIC®のテスト項目（200項目））

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	80.9	136.0	−.58	.22	1.00	.0	.99	.0
S.D.	29.7	.0	1.31	.10	.07	1.1	.15	1.1
MAX.	135.0	136.0	2.87	1.01	1.30	5.1	1.51	4.6
MIN.	9.0	136.0	−5.10	.18	.84	−2.6	.55	−2.3

REAL RMSE	.25	TRUE SD	1.28	SEPARATION	5.13	ITEM RELIABILITY .96
MODEL RMSE	.25	TRUE SD	1.28	SEPARATION	5.18	ITEM RELIABILITY .96
S.E. OF ITEM MEAN = .09						

表4　項目フィット係数

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
2	134	136	−4.40	.71	.99	.2	.60	−.3	.13	.07	98.5	98.5	問2
6	129	136	−3.10	.39	1.03	.2	1.32	.8	.03	.12	94.9	94.9	問6
9	134	136	−4.40	.71	.99	.2	.55	−.4	.14	.07	98.5	98.5	問9
18	63	136	.15	.18	1.14	2.7	1.16	2.3	.10	.30	54.4	62.9	問18
33	117	136	−1.97	.25	1.10	.6	1.31	1.3	.00	.19	86.0	86.0	問33
38	64	136	.11	.18	1.21	4.0	1.22	3.1	.00	.30	47.1	62.8	問38
43	96	136	−.97	.20	1.07	.9	1.41	2.9	.13	.26	67.6	71.0	問43
46	89	136	−.71	.19	1.01	.2	1.29	2.6	.22	.27	66.9	66.9	問46
50	97	136	−1.01	.20	1.05	.6	1.44	3.1	.14	.25	71.3	71.7	問50
56	31	136	1.33	.21	1.14	1.2	1.33	2.1	.05	.29	77.2	78.2	問56
120	52	136	.51	.19	1.19	2.8	1.25	2.9	.03	.30	59.6	66.1	問120
124	42	136	.87	.19	1.15	1.8	1.23	2.1	.07	.30	66.9	71.2	問124
140	9	136	2.87	.36	1.00	.1	1.51	1.3	.10	.22	94.1	93.5	問140
144	59	136	.28	.18	1.30	5.1	1.35	4.6	−.12	.30	45.6	63.8	問144
158	87	136	−.64	.19	.84	−2.6	.79	−2.3	.50	.27	72.1	65.9	問158
159	73	136	−.18	.18	.88	−2.5	.86	−2.1	.46	.29	72.8	62.5	問159
176	58	136	.31	.18	.89	−2.0	.86	−2.0	.46	.30	69.1	64.1	問176
177	24	136	1.68	.23	1.14	1.0	1.42	2.1	.01	.28	83.8	83.0	問177
178	66	136	.05	.18	1.14	2.7	1.16	2.3	.10	.30	55.1	62.6	問178
188	60	136	.24	.18	1.15	2.6	1.18	2.5	.09	.30	56.6	63.6	問188

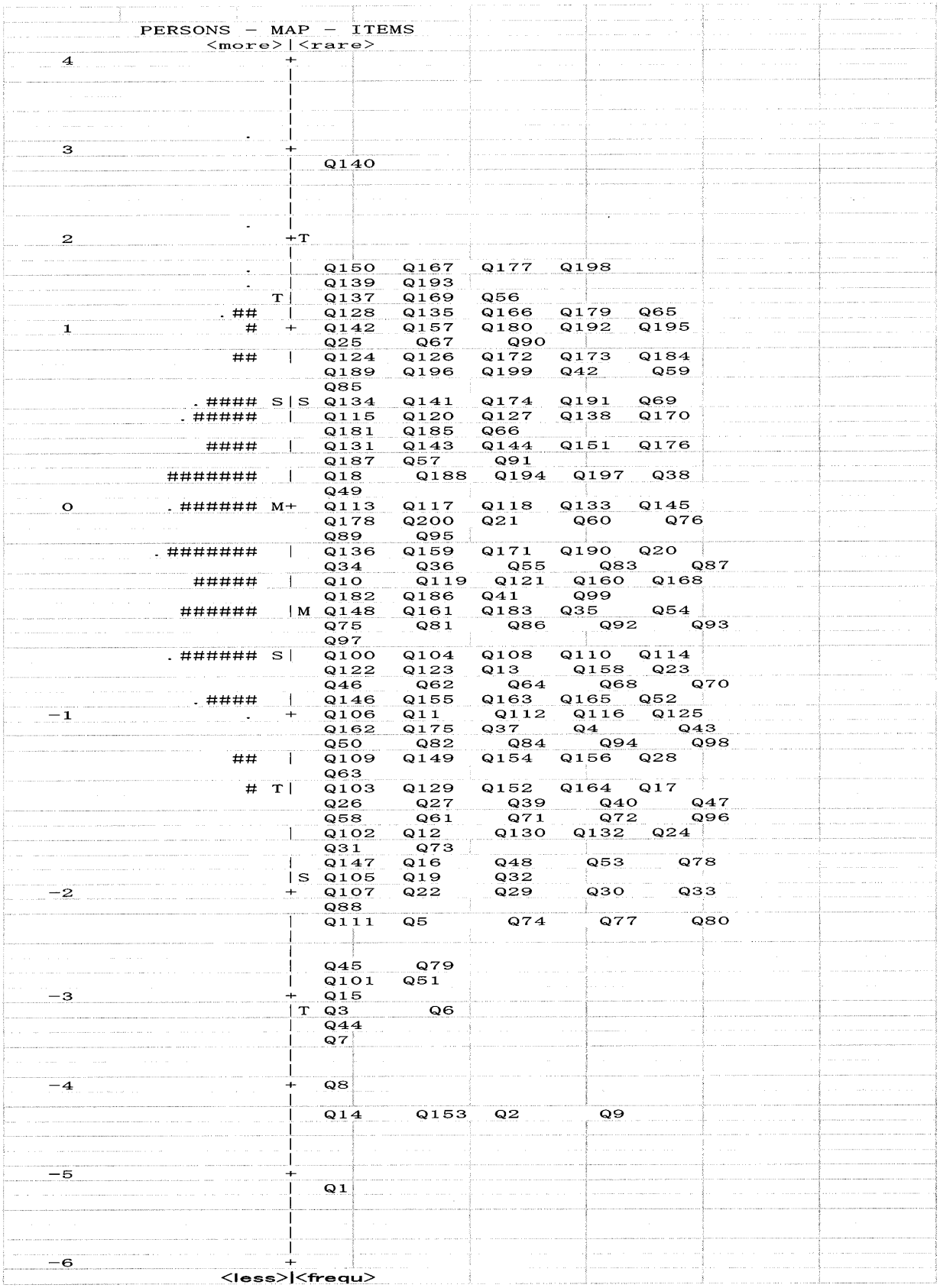

図 1

受験者・項目散布図

インフィットとは能力水準[5]に近い項目に受験者が予想に反して不正解したり、難度水準から離れた受験者が予想に反して正解したりする不適合の度合いを示している。

一方、アウトフィットは能力水準から離れた項目に受験者が予想に反して正解したり、不正解したりする不適合の度合いを示している (Beglar, 2010)。すなわちインフィット値が高い場合、テスト全体が体系的に受験者に対して不適合を起こしている可能性があるのに対し、アウトフィット値が高い場合は正答数が際立って高い、または低い一部の受験者による偶発的で予測に反する解答様式を示している。平方平均値 (MNSQ) と Z 標準化偏差値 (ZSTD) はデータがモデルにどの程度適合したかを示す指標として利用できる。平方平均値は 0.75 から 1.3, Z 標準化偏差値は-2 から+2 までの間であれば許容範囲であると考えられている (McNamara, 1996; Bond & Fox, 2001)。さらに各項目における正答不正答と学習者集団の能力測定値との間の相関係数 (PT-MEASURE CORR. RXP.) が、観察値 (EXT OBS) と期待値 (MATCH EXT) の 2 種類によって示されている。項目の番号は 1 から 100 までは聴解セクション、101 から 200 までは読解セクションの項目を示している。TOEIC®全体の 200 項目中、全体の 1 割に相当する 20 項目がラッシュモデルに適合していなかった。20 項目の内訳を調査すると、聴解セクション、読解セクションそれぞれ 10 項目ずつであった。以下、構成概念的妥当性の各側面について検討を加えていく。

(1) 内容的側面：TOEIC®の各テスト項目は測定対象とする能力を反映しているか。

内容的側面に関わる項目配置に関する情報は表 3 のテスト項目に関する分離係数の値によって判断できる。この値が、5.13 であったことから TOEIC®において適切な項目配置が行われていると判断できる。分離係数を発展させた概念である項目階層値と受験者階層指数は（ (4×分離係数)+1)/3 の公式で求めることができる。この指標について Beglar (2010) はプレイスメントの識別力を示す値として議論している。Beglar (2010) の解釈に従うと研究対象である TOEIC®の項目階層値は、5.7（((4×3.85)+1)/3 =5.47）となり項目難度の階層が 5 つ以上の階層に分かれていることを示している。同様に受験者階層指数は 7.17 になることから受験者は 7 つ以上の能力階層に分れていることを示している。両者ともに項目難度と受験者の能力測定値と観点は異なるが、テストが様々な難度の項目を含み、受験者の能力レベルを識別できる性質の判断材料となる。

これまで述べた結果を図 1 の受験者・項目散布図を利用して確認する。受験者・項目散布図の左側に示された#マークは 2 名の受験者を示し、右側には難度の高い項目番号を上に配置し、下に行くにしたがって難度の低い項目番号が配置されている。研究対象となった TOEIC®の場合、項目番号 Q140 が最も難度の高い項目であり、項目番号 Q1 が最も難度の低い項目であった。この項目難度と受験者の配置を観察すると、英語学力の低い受験者の能力値を推定するための項目は充分に配置されているが、能力値 2 を超える受験者の能力を推定するための項目は、項目番号 Q56, Q65, Q128, Q135, Q137, Q139, Q140, Q150, Q166, Q167, Q169, Q177, Q179, Q193, Q198

の 15 項目である。セクションごとの内訳は聴解が 2 項目、読解が 13 項目である。TOEIC®は 200 項目で構成されているテストである。全体の 7.5 パーセントの項目しか能力値 2 以上の受験者の英語学力の推定のために配置されていないことから、受験者の英語学力が伸長した場合、それを弁別し得点として反映することが充分にできない可能性が考えられる。項目・推定値相関係数を検討したところ、Q14 と Q144 以外がすべて正の相関係数を示していた。TOEIC®全体の中でモデルに適合しない点において不適切と判断された項目の数は 11 であった。したがって内容的側面においては、難度の高い聴解セクションのテスト項目を増やす必要が指摘できるが、それ以外に特別な問題はないと判断できる。

(2) 本質的側面：TOEIC®の各テスト項目への応答パターンはラッシュモデルの予測に適合しているか。

受験者ミスフィット係数を検討したところ、ラッシュモデルに適合しない受験者はいなかった。モデルに適合しない受験者数が全体の 2 パーセント未満であれば応答パターンに問題がなく一定の傾向を示していると考えられる (Pollitt & Hutchinson, 1987)。一方、テストが提示している錯乱肢が受験者のテスト解答行動において機能しているかを示す値である平均能力値 (AVERAGE ABILITY) を調査した。[6] その結果、項目番号 15 と 144 に改善すべき問題点が浮かび上がった。項目 15 は難度の低い項目で受験者 136 名中 128 名（94 パーセント）が正答しており、英語学力の高い受験者と低い受験者の弁別に対して十分な機能を果たしておらず、全体の推定値と有意な相関は認められなかた (r=-.03, *n.s.*)。したがって、項目 15 は弁別機能を果たしていないと判断できる。項目 144 は 136 名の受験者中、59 名（43 パーセント）が正答している。先の項目 15 と同様に、全体の推定値と相関が認められなかったこと(r=-.12, *n.s.*)が問題として挙げられる。以上から受験者全員がラッシュモデルに適合し、錯乱肢分析によって問題点が指摘されたテスト項目は 2 項目に留まったことから、本質的側面に問題はないと判断できる。

(3) 構造的側面：TOEIC®は実施された時点でひとつの能力を測定しているか。すなわち、テストの一次元性が確認できるか。

表 2 と表 3 で示された基本測定値（受験者測定値と項目測定値）における受験者フィット係数と項目フィット係数の結果から、TOEIC®の受験者の解答パターンと項目の難度はラッシュモデルに適合していることが示された。テストの一次元性を前提するラッシュモデルに適合していることから TOEIC®の一次元性が認められると間接的に判断することも可能だが、これらのフィット係数の値は厳密性に欠けるとの指摘もある（例：Smith & Plackner, 2009）。現在ではラッシュ分析で得られた測定値（受験者測定値と項目測定値）がデータの分散をどの程度説明するかを計算し、その結果によって一次元性に関する判断の確証を得ることが多い。例えば Beglar (2010) はラッシュ分析で得られた測定値による分散説明率が 85.6%であると報告し、Linacre (2008) が定めた 60%基準を超えていることを根拠に研究対象である語彙サイズテストの一

次元性を主張した。この例を参考に本研究データに対しても同一の分析を実施したところ、測定値による分散説明率は24.9%に留まった。分散説明率が低いと判断された場合、残差に対する主成分分析 (principle component analysis on the residuals)を行う手法が開発され実施されている (例 : Goh & Aryadoust, 2010)。[7]その結果、残差において多次元性を示唆する複数の contrast の存在が認められた。Linacre (2009) によれば、contrast の値が 2.00 を超えた場合、多次元構造を示唆するとしている。本論文のデータでは、全ての contrast の値が基準値を上回った（4.3~5.7）。分散説明率と残差に対する主成分分析の結果が、受験者フィット係数と項目フィット係数から導き出される結果と対立したことから、現時点ではTOEIC®の構造的側面について明確な結論を導き出すことはできないが、一次元性を強く主張することはできないと考えられる。

(4) 一般化可能性：TOEIC®のテストの得点によって行われた判断は、テストが実施された時点や場所以外の状況でも一般化できる可能性を示しているか。

一般化可能性は実施されたテストの得点とその解釈によって得られた判断が、テストが実施された時点や場所以外の状況でも一般化できるかどうかに関係している。しかし、テストを複数回、同一の受験者集団や異なる受験集団に実施することは難しいことが多い。[8] 本研究では古典的テスト理論でテストの信頼度係数を測定する際に利用される折半法 (split-haif method) の手法を一部応用した (Baghaei, 2008, 2009)。この方法では、研究対象のテストをテスト項目の番号に応じて、奇数番号項目と偶数番号項目に分割し、2 つのテストレットを作成する。すなわち本研究の調査参加者はTOEIC®を 1 度しか受験していないが、その受験時に同時に半分の量の TOEIC®（奇数項目番号項目 TOEIC®テストレットと偶数項目番号 TOEIC®テストレット）を 2 回受験したと想定する。そして 2 つのテストレット間の関係を調査すれば、TOEIC®全体の安定性を検証できる。各テストレットのデータをラッシュモデルによって分析し、それぞれの受験者能力測定値間の類似性を検討した。図 2 はこの手法によって得られた 2 つの受験者能力値を同一受験者ごとにクロス・プロットしたものである。品質管理 (quality control) の概念を援用し、中心線から 95 パーセントの品質管理が保障されるレベルを想定し、管理限界（上方限界と下方限界）を設定した (Bond & Fox, 2007, p. 72; Bond, 2003, p. 187)。それらの限界内に収まれば、2 つの能力測定値は安定状態あると判断することにした。その結果、受験番号 11，17，18，73，132 の 5 名の受験者が限界線から逸脱した。136 名中の受験者の中での 5 名（全体 3.68%）が逸脱している状況から、テストの改善の余地が残されていることは否定できないが、充分な一般化可能性を示していると判断できる。[9]

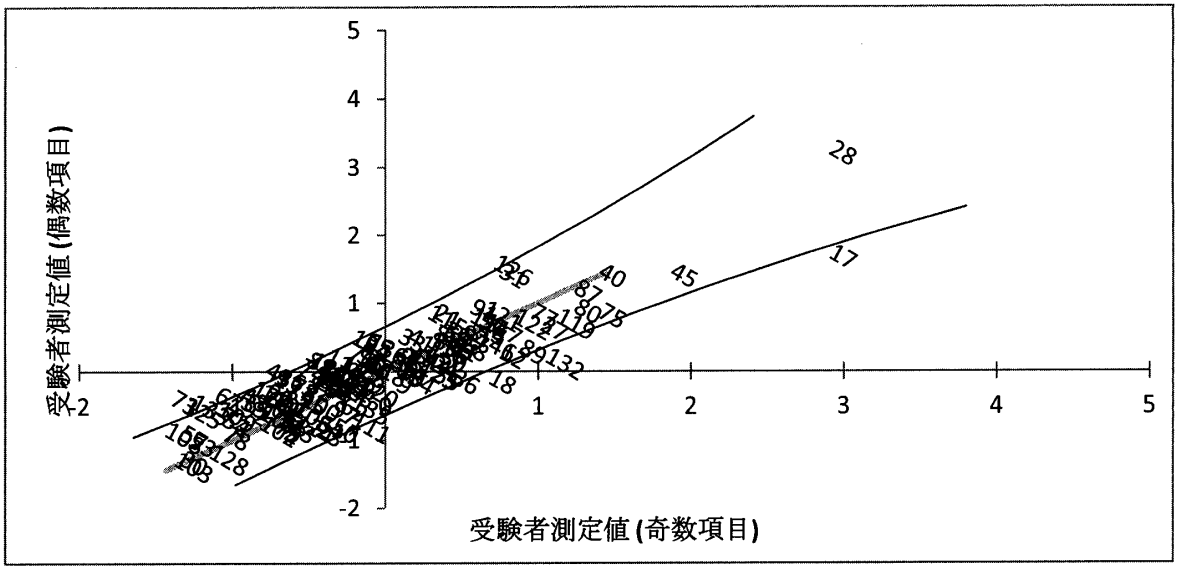

図 2　受験者測定値（奇数項目・偶数項目）クロス・プロット

(5) 外的側面：TOEIC®は、受験者の英語能力が変化した場合、その変化を得点として表す性質を有しているか。

(6) 結果的側面：TOEIC®は、社会の中で行われる意思決定や価値判断に対して利用された場合、多くの人が納得できる結果をもたらす可能性があるか。

外的側面および結果的側面は図 1 の受験者・項目散布図によって検討できる。受験者・項目散布図を検討すると、最も高い能力測定値を示した受験者集団が今後の学習によって英語学力を伸長させた場合、その能力を推定するための項目は 15 項目しか配置されていないことが分かる。TOEIC®の利用法に関係する問題であるが、大学生や成人の英語学習者を念頭においた場合、英語学力の伸長の度合いを測定するため、複数回の受験は一般的に考えられる。能力測定値の変化を念頭に置かないテスト、例えばプレイスメントテストや入学試験であれば、受験者の能力よりも高い難度の項目を意図的に配置する必要はないかもしれない。しかし、複数回の受験がありうる標準英語テストであるならば、より難度の高い項目の構築と配置が必要になるだろう。

一方、この受験者・項目散布図から結果的側面について判断を加えてみる。能力値 -5 から-2 にかけて、ところどころテスト項目が配置されていない個所が散見されるが、受験者の能力値付近にはテスト項目が配置されていない個所はみられない。したがって受験者の能力測定値に適切な難度の項目が配置された状況で英語学力を推定されているため、ラッシュモデルによって得られた能力推定値が社会における意思決定や判断のために有益な情報を提供できる可能性は高いと考えられる。

結論

本研究ではラッシュモデルの分析結果を Messick (1989) が提唱した妥当性の 6 つの側面に結び付け、TOEIC®の構成概念的妥当性を検証した。その結果、内容的側面、本質的側面、一般化可能性、外的側面、結果的側面には問題はなかったが、外的側面

と結果的側面を検討する中で、難度の高いテスト項目を増やす必要性が示唆された。一方、構造的側面の検討において受験者フィット係数および項目フィット係数の結果と、測定値（受験者能力測定値と項目測定値）による分散説明率と残差に対する主成分分析の結果が対立することとなったため、TOEIC®の一次元性について明確な判断を下すことはできなかった。[10] ラッシュモデルに基づく分析結果を踏まえた判断には、拠り所となる絶対的な基準は存在しないため、今後は同種の研究結果が出た場合、どのような解釈が可能か追究していく必要がある。[11] この点に関する研究が進展することで、テストの構成概念的妥当性の検証法としてのラッシュモデルの利用範囲が拡大すると考えられる。

謝辞

本研究のデータ収集に際し、西南学院大学、熊本大学、静岡産業大学、長崎純心大学、東京情報大学の学生および関係者の方々、そして鈴木千鶴子先生（長崎純心大学）に多大なご協力を賜った。財団法人国際ビジネスコミュニケーション協会研究開発部長の三橋峰夫氏には研究計画の立案と調査実施に際し、様々な助言を賜った。本データを利用した研究プロジェクトにおける共同研究者、島谷浩氏（熊本大学）と法月健氏（静岡産業大学）には多くの知的刺激と激励をいただいた。これらの方々に感謝の意を表明したい。本論文は西南学院大学在外研究規定による連合王国セントラル・ランカッシャー大学 (UCLan) における研究期間（平成 24 年 8 月 8 日~平成 25 年 8 月 9 日）において執筆された。本研究は財団法人国際ビジネスコミュニケーション協会の TOEIC®研究助成金（研究題目「新旧 TOEIC®の比較とプレイスメントテストとしての妥当性」（研究代表者：伊藤彰浩）（研究実施期間：2008 年 4 月～2010 年 3 月）により実施されたことも関係者の皆様への感謝を込めて付記しておきたい。

註

1. 近年では Messick (1989) によって妥当性の概念が拡大されたため、テストの妥当性の検証が多様かつ複雑になりすぎているとし、科学哲学者や心理測定学者の一部から妥当性の概念はテストが測定対象とする知識や能力を測定しているか否かという単純な命題に限定すべきとの意見もある。例えば、2004 年以降の Denny Borsboom (Universiteit van Amsterdam) らによる論考 (例：Borsboom & Mellenbergh, 2004)は Messick (1989, 1995) への批判に留まらず、科学哲学にまで及ぶ広範なテストの妥当性に関する論考として注目を浴びている。テストの妥当性の概念の縮小化および単純化の必要性に関する議論については、言語テスト研究の枠組みにおいて論じてみたいトピックである。
2. ラッシュ分析のソフトウェアの中で初心者にも利用しやすいものとしては *Test Data Analysis Program: T.D.A.P. Ver. 2.02*（大友・中村・秋山、2007）が挙げられるが、本論文において挙げた研究はいずれも *WINSTEP* (Linacre, 2008) を利用していたことから分析には同じソフトを用いることにした。
3. この信頼性係数の値は古典的テスト理論で用いられるクロンバック α などのよう

に、観察された分散における真の得点の比率の観点から内的一貫性を検定したものではない。詳細は Smith, Linacre & Smith (2003) を参照されたい。

4. *WINSTEP* (Linacre, 2008)によって得られた測定値については真正性の観点から、必要以上の加工は避け、可能な限りそのまま図表として掲載した。
5. ここでの「能力」とは、TOEIC®が聴解セクションと読解セクションにおけるテスト項目によって測定を意図している英語に関する総合的な言語熟達度のことである。
6. 本研究で利用した TOEIC®データは項目における不正解を 0、正解を 1 としたゼロイチデータであったため、*WINSTEP* (Linacre, 2008) が提供する 4 つの選択肢（正解の選択肢 1 つと錯乱肢 3 つ）それぞれがどのような機能を果たしているかに関する錯乱肢分析を行うことはできなかった。そこで、正解の選択肢か、不正解の選択肢 3 つのいずれかを選んだかにのみ着眼して錯乱肢分析を行い能力平均値を求め、受験者の正答、不正答における解答パターンがどの程度モデルに適合しているかを調査し、本質的側面を検討した。
7. テストを実施した際、受験者は様々な認知的、情緒的な影響を受けていると考えられる (Henning, 1987, p. 142)。したがって、テストが実施された際、そのテストが測定を意図した能力特性のみを測定している可能性は、現実問題として、極めて低いと言わざるを得ない。ラッシュモデルを利用する際のテストの一次元性とは「テスト・データにおけるパターンの単一性」(McNamara, 1996, p. 271) としてとらえる方が現実的かつ実用的であると考えられる。
8. Messick (1989) の妥当性の定義に忠実に従い、一般化可能性を調査するのであれば、英語学力の変化が認められないと考えられる短期間の間に、2 つ以上の TOEIC®を実施し、2 つのテスト得点の間に有意差がないことを確認する方法が適切である。事実、Messick (1989, 1995) の一般化可能性の概念は学習者集団の能力の変化をテストが正しく測定できるかどうかに関係している。
9. 本分析のように単一の受験者集団のデータを奇数、偶数のテスト項目番号に分けた場合、奇数群と偶数群の項目難易度平均が一致しているとは断定できないため、奇数、偶数両方の項目群を受験している受験者能力の平均はそれぞれ 0 に設定される。したがって、平均値が一致するため、ラッシュ能力測定値の平均値の有意差検は必然的に有意差がないことになる。個別の受験者の測定値比較においては *t* 検定の結果、有意な残差が生じていることは、本研究の基本データにおいて明らかになっているが、ラッシュモデルを利用した分析において、平均値を同一尺度に位置づける理念からして、意味のある有意差検定はできない。
10. ラッシュモデルを利用した研究では、当該テストが信頼でき妥当である、という様な説明や判断は適さない。理由は規範的な統計モデルを利用して能力値を予測するため、テスト自体ではなく、そのデータから得られる測定値が信頼性に関係し、特定の目的のために、項目測定値、受験者測定値、そしてフィット係数から得られる情報に基づく推論 (inference) が妥当性に関連するからである (Smith, Linacre & Smith, 2003, pp. 198-204)。

11. 現在、*WINSTEP* 開発者である Linacre 氏の助言を仰ぐとともに、関係する論文の内容を精査している。

参考文献

Anastasi, A. (1954). *Psychological testing.* New York: Macmillan.

American Psychological Association. (1954). *Standards for educational and psychological tests and manuals.* Washington, D.C.: American Psychological Association.

American Psychological Association. (1966). *Standards for educational and psychological tests and manuals.* Washington, D.C.: American Psychological Association.

American Psychological Association. (1974). *Standards for educational and psychological tests and manuals.* Washington, D.C.: American Psychological Association.

Baghaei, P. (2008). The Rasch model as a construct validation tool. *Rasch Measurement Transaction, 22 (1),* 1145-1146.

Baghaei, P. (2009). *Understanding Rasch model.* Mashad: Mashad Islamic Azad University Press.

Baghaei, P. (2010). Validation of a multiple choice English vocabulary test with the Rasch model. *Journal of Language Teaching and Research, 2 (5),* 1052-1060.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing, 27 (1),* 101-118.

Bond, T. G. (2003). Validity and assessment: A Rasch measurement perspective. *Metodologia de las Ciencias del Comportamiento, 5 (2),* 179-194.

Bond, T. G. & Fox, C. M. (2001). *Appying the Rasch model: Fundamental measurement in the human sciences.* Boston, MA: Lawrence Erlbaum Associates.

Borsboom, D. & Mellenbergh, G. J. (2004). The concept of validity. *Psychological Review, 111 (4),* 1061-1071.

Campbell, D. T. & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56(2),* 81-105.

Cronbach, L. J. & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Goh, C. & Aryadoust, V. (2010). Investigating the construct validity of the MELAB listening test through the Rasch analysis and correlated uniqueness modeling. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 8,* 31-68.

Guttman, L. (1950). The basis for scalogram analysis. In Stouffer, S. A., Guttman, L. A., & Schuman, E. A, (Eds.), *Measurement and prediction* (pp. 60-90). Princeton: Princeton University Press.

Henning, G. (1987). *A guide to language testing: Development, evaluation, research.* Boston, MA: Heinle & Heinle Publishers.

Kelly, T. L. (1927). *Interpretation of educational measurement.* New York: Macmillan.

Landy, F. J. (1986). Stamp collecting *versus* science: Validation as hypothesis testing.

American Psychologist, 41, 1183-1192.

Linacre, M. (2008). *WINSTEP Rasch measurement computer program* (Version 3.66.0). Chicago: Winsteps.com.

Linancre, M. (2009). *A user's guide to WINSTEP.* Chicago: Winsteps.com.

Mahoney, S., Matsuura, H. & Murakami, Y. (2007). *New essential listening for the TOEIC® test.* Tokyo: Kinseido.

McNamara, T. F. (1996). *Measuring second language performance.* New York: Longman.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (pp. 13-103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50 (9),* 741-749.

Norizuki, K., Ito, A. & Shimatani, H. (2011). Exploring item-examinee response characteristics in search of diagnostic functions of TOEIC® tests for university students in Japan. *JLTA Journal, 14,* 1-20.

Pollitt, A. & Hutchinson, C. (1987). Calibrating graded assessments: Rasch partial credit analysis of performance in writing. *Language Testing, 4 (1),* 72-92.

Rasch, G. (1960). *Probablistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Measurement.

Shimatani, H., Norizuki, K., Ito, A. & Kinoshita, M. (2009b). Two versions of TOEIC® as a placement test: A comparative study based on the classical test theory and Rasch model. *English Language Assessment, 3,* 39–56.

Smith, R. H., Linacre, J.M., & Smith, Jr., E.V. (2003). Guidelines for manuscripts. *Journal of Applied Measurement, 4,* 198-204.

Smith, R. M. & Plackner, C. (2009). The family approach to assessing fit in Rasch measurement. *Journal of Applied Measurement, 10 (2),* 424-437.

阿部真理子・ウィスナー, ブライアン・酒井英樹 (2008). 「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」『高崎経済大学論集』50 (3・4), 125-134.

伊藤彰浩・島谷浩・法月健・木下正義 (2009a). 「古典的テスト理論による 新・旧 TOEIC®の比較分析」『日本言語テスト学会研究紀要』12, 26-45.

伊藤彰浩・島谷浩・法月健・木下正義 (2009b).「新旧 TOEIC®の受験者配置における平行度の検討と内的妥当性検証」『JACET 九州・沖縄支部東アジア英語教育研究会紀要』3, 5-17.

伊藤彰浩・島谷浩・法月健 (2010).「新・旧 TOEIC®の予測的妥当性の検証」『日本言語テスト学会研究紀要』13, 111-125.

大友賢二 (1996). 『項目応答理論入門』東京：大修館書店.

大友賢二（編）・中村洋一（著）(2002). 『テストで言語能力は測れるか～言語テストデータ分析入門～』東京：桐原書店.

大友賢二・中村洋一・秋山實　(2007). *Test data analysis program: T.D.A.P. Ver. 2.02* [Windows 版] オープンソース・プログラム.

島谷浩・法月健・伊藤彰浩・木下正義 (2009a). 「ラッシュモデルによる新旧 TOEIC® の比較」『JACET 九州・沖縄支部研究紀要』14, 17-32.

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.

TOEIC®運営委員会 (2007).『TOEIC®テスト新公式問題集』東京：TOEIC®運営委員会.