

<書評>

野口裕之・大隅敦子 2014. 『テスティングの基礎理論』
東京：研究社 190 頁

井澤廣行*

Hiroyuki Izawa

評者において、以下に挙げる三点の関連性により、時宜にかなって出版された本書は新鮮であり斬新である。

- 1) 『英語学力の経年変化に関する研究 — 項目応答理論を用いた事後的等化法による共通尺度化』(齊田¹⁾、2014)。
- 2) 2010 年改訂以降の年 2 回実施「日本語能力試験」各級得点表示において、各級得点を異なる受験時の間で比較可能とする項目応答理論適用による等化尺度の採用(本書、p. 9)。
- 3) CEFR(Common European Framework of Reference for Languages: Learning, teaching, assessment)における第 2 言語能力参照枠組みに対して、「日本語能力試験」各級対応位置付けに問題となり得る第 2 言語としての日本語の習得と能力測定に係る独自的特質(本書、第 11 章及び第 12 章)。

共著者である野口は、日本における項目応答理論研究分野での大家の一人であり(野口²⁾、1991; 渡辺・野口³⁾、1999、参照)、上記 2)の実現に寄与した中心人物の一人でもある(本書、参照)。上記 1)との関連は、この書籍の元である齊田による 2010 年提出博士論文の主査が野口であり、(齊田¹⁾、2014, p. 139)、この等化法が本書にて参照されている(pp. 121-122)。共著者である大隅の専門分野の一つは「欧洲における言語テスト開発事情」(本書、p. 190)とあり、上記 3)に結実されている。従って、「日本語能力試験」を中心とする大規模言語テストと項目応答理論のそれぞれを縦糸と横糸として、横糸への導線としての古典的テスト理論を含む本書記述内容である。正答項目数得点に基づく古典的テスト理論における妥当性と信頼性への言及は、代表的な幅広い先行文献への参照で以て小気味よくまとめられており、特に、信頼性係数についての記述は要領を得ている。以下に与えるその縦糸、導線、横糸をつづる章立ては明瞭であり、本書内容の斬新性が示されている。

*流通科学大学人間社会学部、〒 651-2188 神戸市西区学園西町 3-1

(2015 年 3 月 20 日受理)
© 2015 UMDS Research Association

第 1 章	世界の大規模言語テスト
第 2 章	大規模言語テストにおけるテスティングの検討
第 3 章	大規模テスト開発の流れ
第 4 章	テスト項目の分析
第 5 章	テストの妥当性の検討
第 6 章	テストの信頼性の検討
第 7 章	項目応答理論
第 8 章	尺度得点の等化と垂直尺度化
第 9 章	特異項目機能の検出
第 10 章	パフォーマンス測定に関する分析
第 11 章	CEFR と言語テスト
第 12 章	日本語能力測定に関する独自性について

上記の章立てにおいて、項目応答理論とその適用法が第 7 章から第 10 章にかけて概説されている。その概説は、順に、「2 パラメタ・ロジスティック・モデル」(以降、2PLM)、「3 パラメタ・ロジスティック・モデル」(以降、3PLM)、「段階応答モデル」、「単純ラッシュ・モデル」、「拡張ラッシュ・モデルとしての部分得点モデルと評定尺度モデル」、及び、「多相ラッシュ・モデル」についてである。項目応答理論は、本書評冒頭での記載事項 1)と 2)を含めて、現代における『テスティングの基礎理論』を構成するとの著者の視点が明確に示されている。なお、第 9 章において言及されている「Mantel-Haenszel 法」、並びに、第 10 章での「 κ 係数」、「重み付き κ 係数」、「級内相関係数」、及び、「 α 係数」は、素点に基づいて算出しないしは出力される指標である。

本書での著者視点による項目応答理論適用の正答項目数得点分析に優る有用性が 2 箇所(p. 75、及び、pp. 111-112)において記されている。前者を全文引用、後者を要約引用として、以下に通し番号を付して、参照する。

- 1) 「項目の困難度が受験者集団とは独立に定義される」(p. 75)。
- 2) 「受験者の能力(特性)値が解答(回答)した項目群とは独立に定義され、異なるテストを受験した受験者間で結果を[等化された尺度値により]比較できる」(p. 75)。
- 3) 「項目の困難度と受験者の能力(特性)値とが同一の尺度上で表現されるため、当該受験者に呈示した項目が適切であったか否かの判断が容易になる」(p. 75)。
- 4) 「この尺度は間隔尺度水準にあって、変化の度合いを表わすことができる」(p. 75)。
- 5) 正答項目数得点は厳密には順序尺度の水準にあるが、受験者の潜在特性尺度値は間隔尺度の水準にある故に、後者において計量統計分析の上で妥当性の高いデータが得られる(p. 111)。

- 6) 正答項目数得点分析における通過率(正答率)及び点双列相関係数の指標値は受験者集団に依存する。一方、「項目応答理論では各項目の特性(難易度および識別力)が項目特性曲線のパラメタで表わされるため、[中略]受験者集団によらず項目パラメタの不変性が成り立つ」(p. 111)。
- 7) 古典的テスト理論における素点得点に基づく信頼性係数は受験者群全体に対する平均的な精度を示す。一方、項目応答理論におけるテスト情報量は各受験者に対するテストの測定精度として個人ごとの評価を可能とする(p. 111)。
- 8) 正答項目数得点において異なる項目に回答した受験者間の測定結果を比較することは不可能である。一方、項目応答理論の適用により、回答する項目が受験者間で異なる適応型テストにおいて、同一の潜在特性尺度上の値で測定結果を表示することが可能である(pp. 111-112)。
- 9) 正答項目数得点における項目の新規入れ替えは、標準化の手続きのやり直しを必要とする。一方、項目応答理論はテスト項目の更新と標準化の手続きの分離を可能とする故に、項目応答理論の適用によりテスト項目の更新が容易となる(p. 112)。

なお、著者は、項目応答理論の適用に関して、「受験者数の少ない試験やモデルの仮定が満たされない試験に適用することはできない」(p. 112)と付言している。「モデルの仮定」は「局所独立の仮定」(pp. 78-79)を意味しており、その定義は、「潜在特性尺度値をある値に固定した場合に、テストに含まれる n 項目に対する応答は相互に独立に生ずる」(p. 78)と与えられている。著者によれば、「局所独立の仮定」は「テストの 1 次元性」と同義であり、「テストに含まれる項目群に共通な潜在特性が唯一一つである」(p. 79)ことを含意する。

評者において、本書記述での唯一最大の理解難点は上記参照 6)であり、「項目応答理論では各項目の特性(難易度および識別力)が項目特性曲線のパラメタで表わされるため、[中略]受験者集団によらず項目パラメタの不変性が成り立つ」(p. 111)との記述である。野口・熊谷・大隅(2007)による論述「日本語能力試験における級間共通尺度構成」(p. 182、参照)は 2PLM の適用によるものであったと述べられている(p. 123)。従って、「受験者集団によらず項目パラメタの不変性が成り立つ」(p. 111)との記述における項目パラメタは識別力パラメタを含むとの著者見解であると思われる。

上に参照した野口・熊谷・大隅(2007)による論述における 2001 年度日本語能力試験での 1 級、2 級、3 級、4 級の受験者数は、それぞれ 72,432、58,044、57,469、35,889 とある(p. 172)。この様な大規模テストへの 2PLM の適用による項目困難度パラメタと項目識別力パラメタの安定性・信頼性・再現性の程度に関して評者は不詳である。一方、2PLM の項目特性曲線は下に与える関数であり、項目識別力パラメタは、受験者能力と項目困難度の両パラメタの差に乗算されて、モデル上での正答確率 $P(\theta)$ に関係していることが明白である。

$$P(\theta) = \{1 + \exp[-Da(\theta - b)]\}^{-1}$$

θ : 受験者能力パラメタ

D : 尺度因子であり、 $D=1.7$ のときに θ 全域にわたり正規累積 2PLM との違いが 0.01 以下になることが知られている(村木⁴⁾、2011, p. 45)。

a_j : 項目 j の識別力パラメタ

b_j : 項目 j の困難度パラメタ

$P_j(\theta)$: 能力パラメタ θ を持つ受験者の項目 j への正答確率

上記の点に関して、ラッシュ・モデル(Rasch⁵⁾, 1960, 1980 年再版)の世界への伝播貢献において著名な Benjamin D. Wright により、項目識別力パラメタの各値は受験者群全体としての能力パラメタ値分布に依存しており、項目識別力パラメタ値と受験者能力パラメタ値との分離の程度、すなわち、反復推定出力される項目群識別力パラメタ値の安定性・信頼性・再現性の程度に大きな問題がある(Wright⁶⁾, 1977, p. 220)と指摘されている。従って、項目数、及び、項目群難易度に不備のない大規模テストにおいては、受験者群能力パラメタ値分布の相当に高い程度の正規分布性が観察される筈であるから、この場合における 2PLM 適用上での項目困難度と項目識別力の両パラメタ値についての「不变性」が、上記 6)での著者による記述(p. 111)に関係しているとも推測される。項目応答理論に精通している著者であるが故に、その点に関する論理明快な言及が望まれる。なお、本書において静⁷⁾(2007)が参照されており(p. 104)、2PLM と 3PLM における項目識別力パラメタについての静⁷⁾(2007)によるその最尤推定値出力経緯の上での「論理矛盾」の指摘(pp. 359-361)を著者は認知していると思われる。以上が、2PLM の現実適用に関して評者が抱く唯一の疑念である。2PLM における項目識別力パラメタの存在は、測定尺度の厳密な正当性において「項目パラメタの不变性」を可能とするのかとの疑念である。

本書評冒頭での記載事項 2)と 3)により、近年における日本語能力試験内容自体とその測定法・分析法の質的向上の程が窺われる。その全容概略を俯瞰するために、以下に、「第 2 言語としての日本語能力」、及び、「日本語能力に関する測定法・分析法」に関係して、本書での参考文献として掲げられている日本語記述での研究図書名・論文題目の一覧を以下に与える。これにより、当該分野での本書著者が重要と考えている日本語記述研究業績を即座に把握できる。なお、執筆者姓、(出版年)、『研究図書名』あるいは「研究論文題目」との記述法であり、掲載順序は出版年順である。

芝 (1978) 「語彙理解尺度作成の試み」

芝・野口 (1982) 「語彙理解尺度の研究 I — 追跡データによる等化」

芝編著 (1991) 『項目反応理論 — 基礎と応用』

池田 (1994) 『現代テスト理論』

大友 (1996) 『項目応答理論入門 — 言語テスト・データの新しい分析法』

渡辺・野口編著 (1999) 『組織心理測定論 — 項目反応理論のフロンティア』

- 牧野・中島・山内・萩原・池崎・鎌田・斎藤・伊原（2001）『ACTFL-OPI 入門 — 日本語学習者の「話す力」を客観的に測る』
- 日本語能力試験実施委員会・日本語能力試験企画小委員会監修（2003）『平成 13 年度日本語能力試験分析評価に関する報告書』
- 庄司・野口・金澤・青山・伊東・迫田・春原・廣利・和田（2004）「大規模口頭能力試験における分析的評価の試み」
- 三枝編著（2004）「日本語 Can-do-statements 尺度の開発」
- 吉島・大橋訳編（2004）『外国語教育 II — 外国語の学習、教授、評価のためのヨーロッパ共通参照枠』
- 斎田（2005）「学力低下問題を考える — 学力測定の方法論から」
- 島田・三枝・野口（2006）「日本語 Can-do-statements を利用した言語行動記述の試み — 日本語能力試験受験者を対象として」
- 村木（2006）「全米学力調査(NAEP)概説 — テストデザインと統計手法について」
- 井上・孫・野口・酒井（2007）「日本語プレースメントテストにおける DIF 研究」
- 静（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も非なる測定のパラダイム』
- 野口・熊谷・大隅（2007）「日本語能力試験における級間共通尺度構成の試み」
- 野口・熊谷・脇田・和田（2007）「日本語 Can-do-statements における DIF 項目の検出」
- 平井（2007）「主観的評定における評定基準、評定者数、課題数の効果について — 一般化可能性理論による定量的研究」
- 熊谷（2009）「初学者向けの項目反応理論分析プログラム Easy Estimation シリーズの開発」
- 近藤・小森編（2011）『研究社日本語教育事典』
- 澤木（2011）「大規模言語テストの妥当性・有用性検討に関する近年の動向」
- CEFR-J 研究開発チーム（2012）『新しい英語能力到達度指標 CEFR-J 公開シンポジウム予稿集』
- 熊谷（2012）「統合的 DIF 検出法の提案 — “Easy DIF”的開発」
- 国際交流基金（2012）『JF 日本語教育スタンダード 2010 第二版』
- 中津原（2013）「能力基準としての Can-do statements とテストの妥当性を検証する「社会的・認知的枠組み」(Socio-cognitive Framework)について」

2000 年以降での当該研究分野における鍵概念は、「話す力の客観的測定」、「分析的評価」、「日本語 Can-do-statements 尺度」、「ヨーロッパ共通参照枠」、「特異項目機能 DIF」、「日本語能力試験における級間共通尺度構成」、「主観的評定」、並びに、「大規模テストの妥当性と有用性」であることが窺わ

れる。これらはすべて本書での第5章、及び、第8章から第11章にかけて詳述されている。又、2000年までの項目応答理論研究における本書の共著者野口を含む日本での先達によるその理論考察と理論適用萌芽が、熊谷(2009)による「項目反応理論分析プログラム Easy Estimation シリーズの開発」に併せて、本書評冒頭記載事項1)、並びに、2) 2010年以降での日本語能力試験における項目応答理論適用による等化尺度の採用(本書、p.9)に具現されている。従って、テスティングに関する理論と実践の両面において、本書は現代性に適う好著であると評価される。更に、本書は、言語テスト関係者のみならず、「試験にかかるすべての人に」(池田⁸⁾、1992)に対して、「未来のテストへ向けて」(池田⁸⁾、1992, pp. 190-217)、「一斉テストから個別テストへ」(池田⁸⁾、1992, pp. 210-212)、及び、「応答型テストから表出型テストへ」(池田⁸⁾、1992, pp. 213-214)の日本での実現に方法論的示唆を与えており、「日本語能力試験」がその技術的な突破口になり得ると期待される。

最後に、付記として、本書評冒頭記載事項3)に関する研究成果を以下に参照する。それは、第2言語としての日本語の習得に係る独自的特質の一端を示唆するものであり、門外漢にも興味をそそる。その出典は次のもの(本書、p.185、参照)である。

Noguchi, H., Kumagai, R., Osumi, A., & Wakita, T. 2008. Comparing factor structures of the Japanese Language Proficiency Test: Differences in factor structure with increasing language proficiency by native language. 15th World Congress of Applied Linguistics (AILA 2008), Essen, Germany.

著者によるその説明(pp. 174-177)の要旨は以下の通りである。

2001年度、2002年度、並びに、2003年度の日本語能力試験における1級、2級、3級、4級の各受験者群を対象として、「文字・語彙」、「聴解」、「読解・文法」の3類から成るテスト項目群への応答データを因子分析にかけた。年度間での因子分析結果に差異はほとんどなかった。例えば、2003年度の1級受験者群84,550名応答データの因子分析結果において、因子数は2と決定された。その2因子構造は、1)「漢字がもたらす情報を検索処理する能力」を表わす「漢字情報処理因子」、及び、2)「文脈を活用して理解を構築する能力」を表わす「文脈情報処理因子」と解釈された。

更に、2003年度の日本語能力試験における1級、2級、3級、4級の各受験者群を「中国語」、「韓国語」、及び、「その他」(非漢字圏)の3つの母語グループに分けて、各級の3グループ別による受験群応答データを因子分析にかけた。その結果による因子構造の解釈は、表12.4(p.177)に与えるものであった。

「表12.4 2003年度日本語能力試験母語グループ別因子数」(本書p.177での提示表の複製)

レベル	中国語	韓国語	その他
1級	1	2	2
2級	1	1	2
3級	1	1	2
4級	1	1	1

留意として、異なるデータ間においては、因子数が同一であるとしても因子の意味する内容が異なっている可能性もある。その留意の上で、表 12.4 における 2 因子は、上に述べた「漢字情報処理因子」と「文脈情報処理因子」と解釈された。「中国語」グループは、母語での漢字使用により、4 級から 1 級まで漢字情報と文脈情報を区別することなく日本語学習を進め得ると理解された。「韓国語」グループは、現在においては母語で漢字を使用する場面が限定されている。しかし、韓国語に漢字由来の語彙も少なくない故に、2 級・3 級・4 級受験者群においては漢字情報処理が因子として抽出されていないと推測された。「その他」(非漢字圏)グループは、4 級受験者群についてはテストに漢字語彙が少ないとから 1 因子であり、3 級以上の受験者群に関しては、母語とは異なる表意文字としての「漢字」の習得及びかな交じり文の読解において、「文脈情報処理」とは異なる「漢字情報処理」の能力が必要になると推察された(以上、本書 pp. 174-177 における著者説明の要旨)。

この考察で以て、著者は、「日本語の学習基準や日本語能力を測定するテストの結果を CEFR など欧米系の言語を念頭において開発された言語能力基準や参照枠に関係づける際には、欧州系の言語にはない漢字情報処理因子に注意する必要がある」(pp. 177-178)と結語している。上記の研究考察に関する評者の興味一端は、「韓国語」と「その他」のグループにおける日本語学習者が、1 級を超えるどの時点で、どの様に、漢字情報と文脈情報を一体化するのであろうかということである。

【参考文献】

- 1) 斎田智里 2014. 『英語学力の経年変化に関する研究 — 項目応答理論を用いた事後の等化法による共通尺度化』 東京： 風間書房
- 2) 野口裕之 1991. 3 章「項目反応理論にもとづくテストの作成法」、芝祐順編 『項目反応理論 — 基礎と応用』(pp. 51-86) 東京： 東京大学出版会
- 3) 渡辺直登・野口裕之編著 1999. 『組織心理測定論 — 項目反応理論のフロンティア』 東京： 白桃書房
- 4) 村木英治 2011. 『項目反応理論』 東京： 朝倉書店
- 5) G. Rasch. 1960. *Probabilistic models for some intelligence and attainment tests*. The Danish Institute for Educational Research. (Reprinted in 1980 by the University of Chicago Press with a Foreword and Afterword by B. D. Wright.)
- 6) B. D. Wright. 1977. Misunderstanding the Rasch model. *Journal of Educational Measurement*, 14, 3, pp. 219-225.
- 7) 静 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
- 8) 池田 央 1992. 『テストの科学 — 試験にかかるすべての人に』 東京： 日本文化科学社