特集論文

外国語としての日本語能力測定を支えるテスト理論

野口裕之*， 倉元直樹**

*名古屋大学大学院教育発達科学研究科 **東北大学高等教育開発推進センター

Test Theory sustaining the Japanese Language Proficiency Test for Non-native Speakers

Hiroyuki Noguchi*, Naoki T. Kuramoto**

* Graduate School of Education and Human Development, Nagoya University

** Center for the Advancement of Higher Education, Tohoku University

The Japanese Language Proficiency Test has been growing as on of the largest language testing program for non-native speakers in the world, following two large scale English testing programs. It has more than half a million applicants per year in recent years. In the present paper, we reviewed theoretical and technical bases called Test Theories sustaining large scale testing programs including language test such as JLPT. Although we have CTT and IRT as two major test theories, IRT has some advantages over CTT on large scale testing. While CTT scores consist of the weighted count of correct answers, IRT provides mathematical models for the scale on latent traits. Since ability parameters can be calibrated independently of item parameter distribution, we can directly compare examination results obtained from different item sets. However we need a computer and an estimation program to calculate test results. We can equate different scales after calibration utilizing anchor items or common test takers. We can also check whether the test items are working fair enough to different subgroups of people with Differential Item Functioning methods. According to DIF analysis, most of the JLPT items were proved to be working unbiased between Korean and Chinese native speakers. JLPT is now preparing for references to interpret scores making use of the Japanese can-do-statements. In order to meet international standards for testing, JLPT has to overcome some barriers brought from customs on examinations based on traditions in the Japanese society.

Keywords：The Japanese Language Proficiency Test, Test Theory, Item Response Theory Test equating, Japanese Can-do statements

キーワード： 日本語能力試験，テスト理論，項目応答理論，等化，日本語 Can-do-statements

1．はじめに

わが国では，最近外国語能力を測定する公的な試験が大学生などを中心に普及し，受験者数が増加している。北米地域の大学・大学院へ留学する際に受験する TOEFL，英国や豪州を中心にやはり留学に際して英語力を測定する試験

〒464-8601 愛知県名古屋市千種区不老町 名古屋大学大学院教育発達科学研究科

Correspondence concerning this article should be sent to: Hiroyuki Noguchi, Furo-cho, Chikusa-ku, Nagoya, 464-8601, JAPAN Email: c31314a@cc.nagoya-u.ac.jp

である IELTS, 留学に限らないがフランス語の能力を測定する DELF/DALF, ドイツ語ではゲーテ・インスティトゥートのドイツ語検定試験，中国語では漢語水平考試，また，企業内での昇進資格などに利用され日本と韓国を中心として受験者数の多い TOEIC など，さまざまな公的な外国語試験が年間複数回実施されている。特に TOEFL は以前から学術分野で必要な英語能力を測定する試験としての権威を各方面で認められていた。

ただ，TOEFL の英語能力を測定する試験としての認知度の高さに比べて，その質を維持するために背後でどのような努力がなされているのかに関しては，わが国ではあまり認識されて来なかったように思われる。外国語教育の立場から問題構成や内容が十分検討されていることは比較的容易に想像されるが，どのようにして試験の品質の評価が行われ，また，異なる時期に受験しても得点を相互に比較可能にするためにどのような措置がとられているのか，性別や人種などで不公平のない項目であることをどのようにして証明するのか，などいわゆる「テスト理論」に基づく試験の設計・開発・分析・評価に多大な努力が傾注されていることに対してはわが国ではほとんど認識されていない。

良質な試験を開発するためには，人材・時間・予算が必要である。上記の外国語試験ではいずれも政府が国際交流や国際理解のために自国の言語を世界に普及することの重要性を認識して，人材・時間・予算を投入して来た結果であり，現在もなお常に改善し続けている。

わが国の場合，人材という点では，大学・大学院でテスト理論に関する講義や演習が開講されているところはきわめて少なく，人材養成の面で心許ない状況にある。テスト理論の研究者が，具体的なテストや調査に関わってその有用性を示していく必要がある。

本稿では，外国語としての日本語能力の測定に関して，テスト理論がどのように貢献しているか，また，できる可能性があるのかについて展望する。

2. 日本語に関する代表的な公的試験

外国語としての日本語に関する代表的な公的試験は現在 3 種類存在している。

まず，「日本語能力試験」が挙げられるが，この試験は独立行政法人国際交流基金と財団法人日本国際教育支援協会とが共催している。

次に，「BJT ビジネス日本語能力テスト」が挙げられるが，この試験は日本貿易振興機構（JETRO）が平成 8 年（1996 年）から実施している。

さらに，「日本留学試験」が挙げられる。この試験は独立行政法人日本学生

支援機構が平成 14 年（2002 年）から実施しており，日本の高等学校の教科に相当する科目で構成されているが，それに加えて「日本語」が試験科目とされている。

以下，それぞれについて具体的な内容を紹介する。

2．1．日本語能力試験

日本語能力試験は昭和 59 年（1984 年）に開始された日本語に関する試験の嚆矢である。外国語としての日本語学習者に目標としての試験を提供する目的，すなわち，学習奨励を目的で開始された。また，日本留学試験の開始以前には，日本語能力試験の結果が各大学に通知され，日本語能力の判断を各大学に任せる形態が取られていた。したがって，日本語能力試験には日本語の学習奨励と留学生のための入学試験という 2 つの目的が課せられていた。その後，日本留学試験に日本語科目ができたことにより，純粋に日本語の学習者に対する学習奨励を目的とした試験となり，テストとして目的が明確になった。また，入学試験を分離することによって，試験の透明性，公開性を高めることも可能となった。

日本語能力試験は，日本国内及び海外において，原則として日本語を母語としない人を対象として日本語の能力を測定し，認定することを目的として実施される。能力水準に応じて「1 級」から「4 級」の 4 つの「級」が設定されている。4 級は，初級学習者に対するものであり，1 級は上級学習者に対応するものである。

内容領域は 3 つの「類」に分かれている。具体的には，「文字・語彙」，「聴解」，「読解・文法」である。すなわち，「話す試験」と「書く試験」が現段階では含まれていない。日本語の産出能力は測定目的に含まれていないということになる。受験対象者は特に限定されていないが，「日本語を母語としない」ということが条件となる。

なお，検定料がドル建てで決められている TOEFL とは異なり，日本語能力試験は国の実状に応じて受験料が決められている。日本語学習を奨励しようという意図の下で取られている措置である。

平成 20 年（2008 年）においては，受験者数は 4 つの級と国内外の実施地の全てを合わせて年間 595,784 名に上った。大学入試センター試験の英語（筆記）の受験者数が 500,297 名（2009 年本試験受験者） であるので，それを上回る規模の試験であることが分かる。

他の外国語試験の場合，TOEIC が最大で約 3,400,000 名（2004 年） ，TO

EFL が 577,038 名（2004 年），IELTS が 2008 年 6 月に年間総受験者数が 1,000,000 名を超え（http://www.britishcouncil.org/japan-about-us-ielts-press-release-june08），漢語水平考試が 246,977 名（2005 年）などである。

日本語能力試験は開始から 20 年以上が経過し，その間に大量のデータが蓄積されてきた。そのデータを踏まえ，現在の言語教育研究，テスト理論の最新の知見と照らし合わせて，抜本的な改善が行われ始めている。2008 年 5 月に発表された中間報告（日本語能力試験改善に関する検討会・国際交流基金・日本国際教育支援協会，2008）によると，

1）課題遂行能力とそのためのコミュニケーション能力を測定するが，これらの能力を支える基礎となる言語知識についても測定する。
2）学習者の実際の言語行動を反映する試験とする。
3）受験者の日本語能力の多様性に対応できるよう，現行試験の 4 レベルから 5 レベルに，レベル調整を行う。
4）「○○ができる（Can-do Statements）」による参考情報を提示する。
5）世界の大規模言語テストで実施されている得点等化を実施する。

などが予定されている。

2. 2. BJT ビジネス日本語能力テスト

BJT ビジネス日本語能力テストは「日本語を母語としない者」が対象という点では日本語能力試験と同じと言えるが，ビジネス関係者を主な対象者としていることが大きな特徴である。できる限り客観的にさまざまなビジネス場面，状況での日本語によるコミュニケーション能力を測定・評価することを目的としている。平成 20 年（2008 年）に実施された第 18 回テストの受験者数は 4,493 名であった。

このテストは，ジェトロビジネス日本語テストとして平成 8 年（1996 年）に開始されたが，平成 15 年（2003 年）以降に大改定をして，「類別」を意識しない「聴読解」形式となった。その中には「聴解のみ」，「読解のみ」，「聴読解両方」を測定する 3 つのタイプの問題項目があるが，類別とはしていない。なお，聴読解テストの高得点者にはオーラル・コミュニケーション・テストの受験資格が与えられる。オーラル・コミュニケーション・テストにおいては文法的に正確であることも重要だが，最終的に話が通じるかどうかがより重視される。したがって，ビジネス場面に限定された試験である。オーラル・コミュニケーション・テストの評価は，ビジネス・テスターと日本語テスターの双方が協力して行うこととなっている。なお，2009 年度からは実施主体が JETRO

から（財）日本漢字能力検定協会に移管される。

2．3．日本留学試験

日本留学試験は，外国人留学生として日本の大学に留学を希望する者について，日本の大学が必要とする日本語力及び基礎学力の評価を行うことを目的として実施される。すなわち，受験者を日本への留学生に特化していることが最大の特徴である。年間に 2 回実施されるが，平成 20 年（2008 年）に実施された 2 回の試験の合計受験者数は 40,536 名であった。

日本留学試験では「記述」，「読解」，「聴解」，「聴読解」という類別構成となっており，ジェトロビジネス日本語テストに近いが，「聴解」と「聴読解」を分けているという特徴がある。「記述」の採点には技術的な困難が伴い，「0，1，2，3」の 4 段階だが，他の類は得点化されている。なお，日本留学試験には英語は含まれず，受験者の英語能力の評価方法は各大学に任されている。

日本語能力試験が日本語に関する試験の嚆矢として，手探りの状況から構築されていったのに対して，BJT ビジネス日本語能力テストと日本留学試験の科目「日本語」は，日本語能力試験の経験を踏まえ，さらに新しい言語教育理論，テスト理論の新しい流れを取り入れたものを加えていった。

3．テスト理論の基礎

3．1．テスト理論の必要性

テスト理論は Gulliksen（1950）など 1950 年代に基本が確立されていた。日本の公的試験ではテスト理論に基づきテストを開発・評価することの重要性について大きな注意が払われることがなかったが，最近になって多くの試験でテスト理論に基づく分析を必要とすることが大きな流れになっている。例えば，日本留学試験は年に 2 回実施されており，試験結果は 2 年間有効となっている。したがって，4 つの試験結果のうち，利用されるのはどの機会の試験で得られた成績でもよいことになる。さらに，実施地域間の時差等への対策から，同一時期の試験でも複数種類の問題が用意されている。そのうちの一つのバージョンが公開されているが，それ以外は非公開である。問題が異なるテストの得点を相互に比較可能にするためには，得点の等化（equating）が必要となる。ある一時期に実施した複数のバージョンも，複数の実施機会で得られた得点も相互に比較可能でなければならない。それを可能にするのがテスト理論である。それは，決してこの試験に特殊な状況ではない。例えば，TOEFL-PBT の場

合，年間に数回実施されていたが，その時期毎に問題項目は異なっていた。しかしながら，TOEFL-PBT では得点の等化が行われているため，常に同じ尺度上で得点が表され，異なる時期間での得点の比較可能性が保証されている。

質の高い大規模な試験を開発・維持するには，測定内容面からの検討を十分に実施しなければならない。妥当性（validity）の問題である。例えば，学力調査であるならば，学習指導要領や準拠した教科書等で教えられる内容もしくはその結果獲得される能力を偏りなく反映していることが必要である。これらの諸問題を理論的，実際的に解決する方策を導くのがテスト理論の役割ということになる。

3．2．テスト理論の分類

テスト理論は，古典的テスト理論と項目応答理論（Item Response Theory; IRT）に大別される。項目応答理論は項目反応理論とも呼ばれるが言語テスト関係では項目応答理論がよく用いられる。

古典的テスト理論は，基本的には正答項目数を数え上げる計数モデルである。古典的テスト理論には，項目分析（item analysis），信頼性（reliability），妥当性（validity）等，心理学的な測定研究などでよく用いられる重要な概念が含まれている。

項目応答理論は，受験者個人の測定結果がモデルで想定される間隔尺度水準の特性尺度値上の１点に位置づけられる「計量モデル」であり，正答項目数を数え上げるのではなく，モデル上テストが直接何らかの潜在的な特性を測定していると考える。

3．3．正答数得点の問題点

受験者が正答した項目数を数え上げて得点とするのが，正答数得点である。項目の重要度に応じた重みをかけて足し合わせることが多い。通常「配点」と呼ばれるのはその重みのことである。我が国では一般的な成績表示の方法だが，国際標準のテストでは正答数得点が用いられることはない。その理由は，同一の受験者でも問題項目が異なると同じ得点にならず，問題項目の異なるテストを相互に比較できないからである。すなわち，当該受験者の特性（能力）水準を一意的に数値で示すことができないという難点がある。例えば，TOEFL の CBT（Computer Based Test）版（現在、日本では実施されていない）ではコンピュータを相手に解答していくが，一人ひとり他の受験者とは異なる問題に解答することになる。全ての受験者が異なる問題で構成されているテストを受

験するにも関わらず，結果は同じ尺度上に表示される。そのような場合，正答数得点は意味を成さないのである。

それでは，標準得点（standardized score）の場合にこの難点が克服されると言えるであろうか。標準得点というのは，テスト得点から平均点を引いて標準偏差で割って算出する。統計学でいう標準化という操作である。平均得点が 0 になるが，負の数や小数点以下の数値が現れるために扱いにくい。そこで，標準得点を 10 倍して 50 加えれば，得点をほぼ正の整数値で表すことができるということで，「標準得点×10＋50＝Z 得点」と表わす。Z 得点は心理学でよく用いられてきたが，教育の分野では「偏差値」として有名になってしまった。

標準得点化によって，成績は受験者集団におけるテスト得点の分布状況とは無関係に表示できる。すなわち，平均値，標準偏差によらず，個々の受験者を当該受験者集団内での相対的な位置で表わすことができる。同一の受験者集団，もしくは，同一の能力水準や等価な複数の受験者集団が難易度の異なる 2 つのテストを受験した場合でも，2 つのテストの測定結果を標準得点で表示すると相互に比較可能になる。

3．4．共通尺度の必要性

しかしながら，標準得点化を行ったとしても問題点は残る。上記の比較は，同一の受験者集団，もしくは等価な複数の受験者集団が複数のテストを受験した場合にのみに成立するためである。実際にはそのような条件が整うケースは限られており，実用性は低い。また，正答数得点の上限や下限である満点や 0 点の位置が，標準得点化することによって揃わなくなることも問題になる。

それに対してテスト得点の等化（equating）は，それを行うことによって複数の受験者集団が異なるテストを受験した場合でも成績を比較することが可能となる。例えば，日本留学試験の場合，年間 2 回の実施時期の受験者集団の能力水準が同じであると考えれば，標準得点化で概ね解決がつくかもしれないが，現実にはそのことが保証されるわけではない。年に複数回実施される米国のテストでも，時期によって受験者集団の能力水準が異なる場合があることが知られている。能力水準が同等とは限らない複数の受験者集団が，それぞれ異なるテストを受験している場合，正答数得点を出しても相互に比較できない。比較可能にするには，標準得点化をさらに進めて，何らかの方法で共通尺度を構成することが必要になる。異なるテストの結果が共通尺度上の値で表現される必要があると言える。

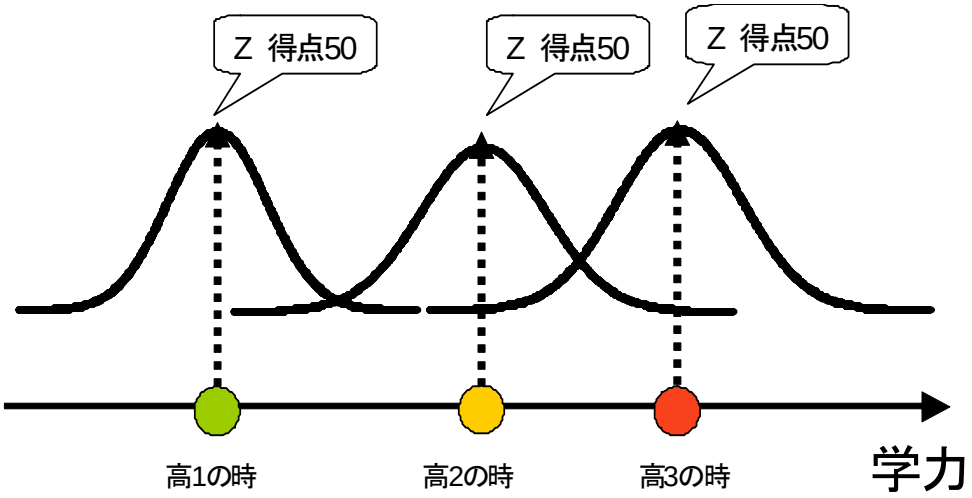

図 1. 共通尺度と学年別標準得点

共通尺度による表現は図 1 のように例示できる。例えば，高 1，高 2，高 3 と Z 得点が Z＝50 と変わらない生徒のケースを考える。標準得点では学力の変化を表すことができない。学力が伸びたにも関わらず，常に相対的には中央に位置した状態である。それを共通尺度上に表すと変化を表すことができる。集団全体では常に中央に位置しているが，集団全体は進歩している。その変化も評価することが必要である。共通尺度には集団自体の変化を表すことが可能という利点もある。

3. 5. 特性尺度値による表現

そこで，後述の項目応答理論につながる簡単なモデルを例示すると以下のようになる。

ⅰ）ある能力を表す潜在特性尺度を一次元の尺度上で表す。

ⅱ）各項目を困難度に応じて直線上の一点に位置づける。

ⅲ）各受験者を能力水準に応じて直線上の一点に位置づける。

ⅳ）各受験者は特性尺度上で自分の位置よりも「左側にある項目」には常に「正答」し，「右側にある項目」には常に「誤答」する。

と仮定する。

共通尺度上の位置が等しい受験者 A と B がそれぞれ，「テスト X」，「テスト

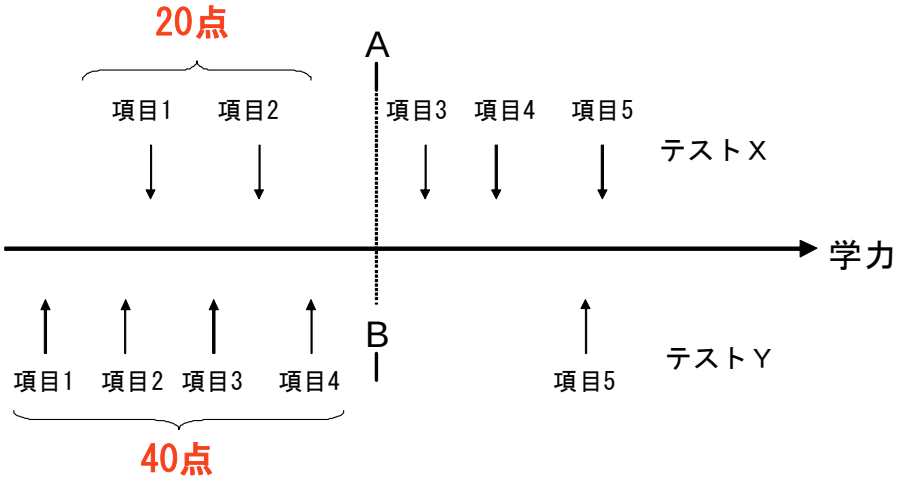

図 2. 学力が等しくても得点が異なるケース

Y」に解答する状況を考える（図 2 参照）。尺度の左側が易しい項目，右側が難しい項目とする。上記の 4 つの仮定の下で，A は 1 問 10 点として 20 点となる。一方，同様に考えると，B は 40 点得ることになる。結果的に，A と B は同じ能力であるにもかかわらず，正答数得点では成績が異なってしまう。

さらに，共通尺度上での受験者間の能力値の変化が等しくとも，テストに含まれる項目の困難度の分布状況によって正答数得点の差が等しくならない状況も考えられる。図 3 のように，仮定された潜在特性尺度上で A は「A1」から「A2」へと能力を伸ばし，B は「B1」から「B2」へと能力を伸ばしたとする。このときの A と B の能力の伸びは等しいとする。A は「A1」の状態では 1 問正答で 10 点，「A2」では 2 問正答で 20 点となる。「A1」と「A2」の正答数得点の差は 10 点である。すなわち，A の能力の伸びは 10 点分と評価されることになる。一方，「B1」は 20 点，「B2」は 50 点となり，その差は 30 点である。B の能力の伸びは 30 点分と評価される。共通尺度上では等しい差が正答数得点では 10 点差，30 点差，と異なってしまうことによって不公平な評価を受けることになる。

これらの諸条件を満たす特性尺度は，項目応答理論を適用すれば実現可能となる。ただし，仮定の「ⅳ）自分の尺度特性値より左側の項目には常に正答し，右側の項目に常に誤答する」，はあまりにも制約が厳しく，非現実的である。項目応答理論では，受験者は特性尺度上で左側にある項目ほど正答する確率が高く，右側にある項目ほど正答する確率が低いという，より緩やかな仮定をおいている。

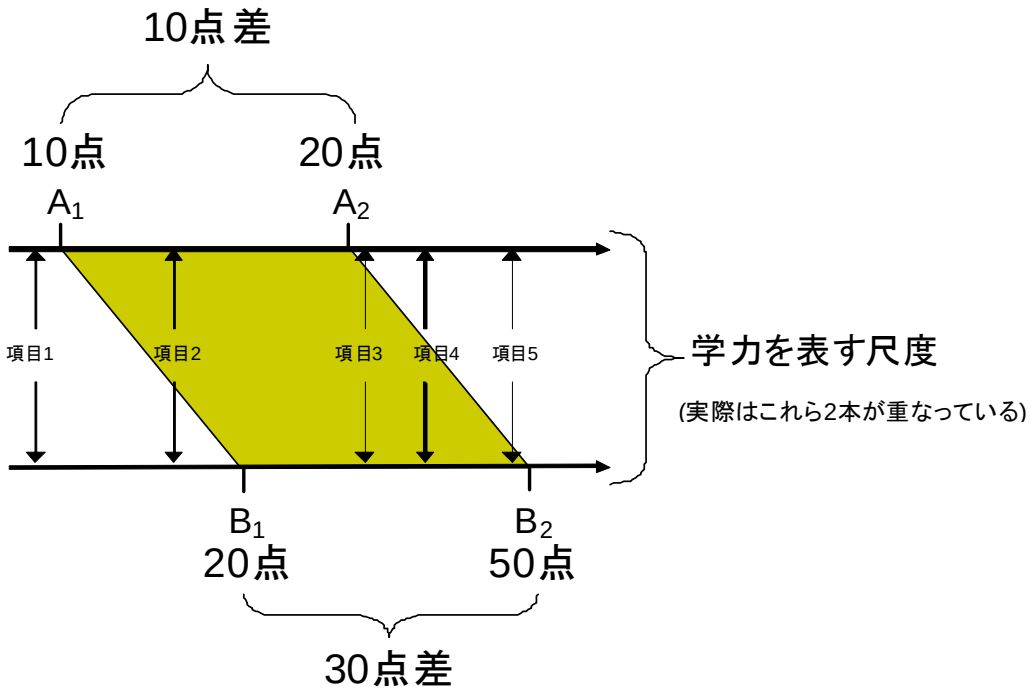

図 3. 等しい能力差でも得点差が異なるケース

4. 項目応答理論

4. 1. 項目応答理論の利点

最初に項目応答理論（以下，IRT と略記する）の利点についてまとめる。

正答数得点による成績表示を前提とする古典的テスト理論の場合には，項目の困難度は正答率（通過率）で定義される。正答率とは，「受験者集団の中で正答した者の全受験者に対する比率」である。したがって，定義そのものの中に「受験者集団で」という限定が含まれる。すなわち，テストの結果は常に当該のテストを一緒に受験した同じ集団の制約の中でしか表現できないのである。一方，IRT では，項目の困難度が受験者集団とは独立に定義される。困難度は IRT モデルで設定される尺度上の位置で定義されるので，受験者集団と全く独立に表わされる。同時に，受験者の特性尺度値も解答する項目と独立に定義され，項目困難度と同様に共通尺度上の一点に位置付けられる。項目困難度と受験者の特性尺度値が同一の尺度上に位置付けて表されることが大きな利点になる。

正答数得点の下では，受験者の成績は n 点満点のテストで x 点という形で表され，項目の困難度は正答率で表されるため，数値そのものだけでは適切な

問題項目が出題されたかという測定の質を検討することができない。IRT は項目と受験者が同一尺度上に位置付くので, 受験者の能力測定に適切な困難度の項目かどうかの判断が容易に可能となるのである。

古典的テスト理論の下では，測定精度は信頼性係数で表わされるが，これは一つのテストに対して一つの数値として定義される。実際には一つのテストが全ての受験者に対して同じ精度を持つことはない。例えば，易しい問題項目から構成された学力テストでは，学力が低い層については精度が高くても，学力が高い層を測るテストとしては有効ではないが, 信頼性係数はそれとは無関係に平均的な値で表される。一方，IRT では測定精度がテスト情報量で表わされるが，これは特性尺度値の関数として定義されるので，尺度値ごとにきめの細かい測定精度の評価が可能となる。

4．2．項目応答理論の成立と発展

IRT をひとつの体系として理論的な開発・整備を行ったのは米国の F. M. Lord である。Lord 以前にも先駆的な研究が存在するが，Lord が 1952 年，Psychometric Society が発行するモノグラフに基本的な枠組みを発表している。当然，その後に理論的発展があるが，基本はそこに発表されている。例えば，最近，構造方程式モデル（SEM: Structural Equation Modeling）を用いた IRT の項目パラメタ値の推定法が提案されているが，基本となる数式は既に Lord（1952）に発表されている。さらに，IRT が統計数理的に確立されたのは Lord & Novick (1968) の出版による。

実用面で IRT が体系的に用いられるようになったのも最近のことではない。TOEFL など米国のテスト開発機関（例えば，ETS や ACT など）で開発される試験を中心に実用水準でも以前から用いられている。一方，わが国の公的試験で用いられるようになったのは最近のことである。

IRT は米国のみならず, 欧州における言語テストや欧州の影響が大きいオーストラリアで，また，医療の分野でも QOL の尺度を構成するなどにも広く用いられている。

4．3．項目応答理論のモデル

IRT では，2 値型の応答モデルが基本となる。受験者の応答が「正答」と「誤答」の二つの段階で表される場合に適用される。2 値型応答モデル以外に，多段階の応答モデルなども存在する。受験者の応答が段階付けられたカテゴリで表わされるモデルである。多枝選択形式の場合は，通常正答と誤答の 2 値型で

採点することが多いので 2 値型応答モデルで十分だが, 選択枝に部分点を付けておくことによって，多値型応答モデルを使うことも可能となる。

2 値型応答モデルで代表的なのは項目特性曲線にロジスティック関数を用いたロジスティック・モデル（Logistic Model）である。後述する正規累積型モデルに比べて理論的な計算が容易だという利点がある。ロジスティック・モデルには，1 パラメタ，2 パラメタ，3 パラメタ・モデルがある。1 パラメタのロジスティック・モデルは, 別名ラッシュ・モデル（Rasch Model）とも言う。デンマークの数学者 Georg Rasch が別の文脈で考案したモデル（例えば，静（2007） などを参照）ではあるが, 数学的には 1 パラメタ・ロジスティック・モデルと同値であることが示される。ラッシュ・モデルは 2 パラメタ，3 パラメタ・ロジスティック・モデルの特別なケースとして位置付けられる。なお, ラッシュ・モデルは欧州やオーストラリアでよく使われている。3 パラメタ・ロジスティック・モデルは米国で多く，2 パラメタ・ロジスティック・モデルは日本で多く見られるという特徴がある。

Lord (1952) では正規累積曲線モデルが提案されていた。心理学では正規累積型曲線がよく用いられるが，それをテスト理論に適用したものである。曲線形は単純だが, 関数型が積分のインテグラル記号の上限に複雑な数式が入る上に，積分式をそのまま用いるので理論的な展開が難しいという難点があった。そこで，正規累積曲線とほとんど変わらない (実用的には等しい) 形状を描くロジスティック曲線が使われるようになった。

2 パラメタ・ロジスティック・モデルでは，特性尺度値 θ を持つ受験者が項目 j に正答する確率は (1) 式で表される。

$$P_j(\theta) = \frac{1}{1 + exp\{-1.7a_j(\theta - b_j)\}} \qquad (1)$$

力が高くなれば正答確率が上がり，能力が低くなれば正答確率が下がる，ということを θ の単調増加関数として表わしている。なお，θ は受験者の特性尺度値を表わす変数，a_j，b_j は項目のパラメタで，曲線型を決定する。分母の指数関数の中の -1.7 という係数は，ロジスティック・モデルを a_j，b_j が等しい正規累積モデルの曲線に近似するために必要な調整値である。

ここで，二つの項目を考える。縦軸が正答確率 $P_j(\theta)$，横軸が潜在特性尺度 θ を表すとする。(1) 式では，特性尺度値 θ が上がるほど正答確率が上がる。正答確率は単調増加関数であり，縦軸の値が逆転することはない。図 4 は (1) 式

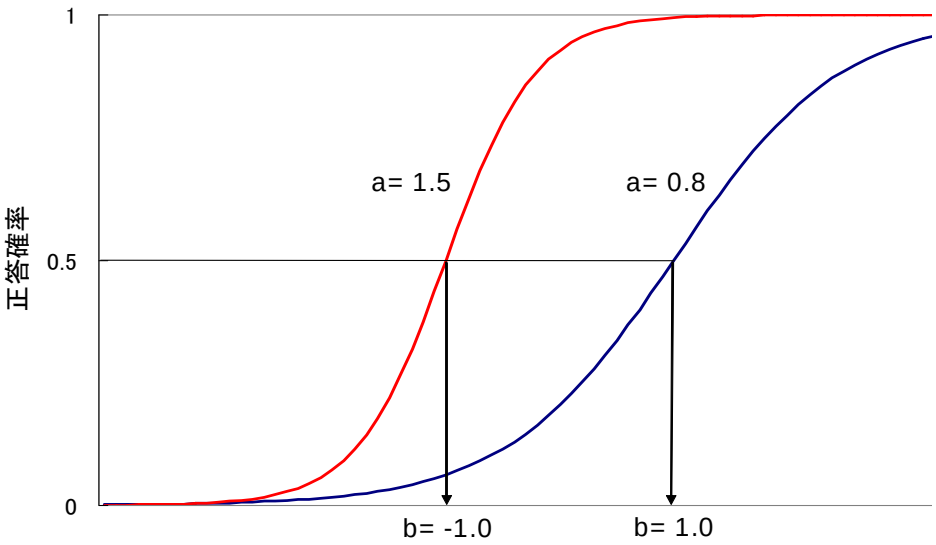

図 4. 2 パラメタ・ロジスティック・モデルの項目特性曲線と項目パラメタ

で表される特性尺度値 θ と正答確率 $P_j(\theta)$ の関数を表す曲線である。項目特性曲線（Item Characteristic Curve）と呼ばれる。図 4 では a_j, b_j の値が異なる 2 つの項目の項目特性曲線を表している。左側の項目では a 値が 1.5，右側が 0.8. b 値は左が -1.0, 右が 1.0 である。b は曲線の位置を表すパラメタである。b 値が大きい項目の項目特性曲線が図の右側に位置する。a 値は曲線の傾きを表す。図 4 では，1.5 の方が 0.8 より急峻なカーブを描いている。

b パラメタは，項目困難度を表す。 θ= -1.0 の能力の受験者が右の項目に正答できる確率は非常に低い。一方，左の項目に正答できる確率は 0.5 となる。b が曲線の位置を表すので，b の値が大きいと難しく，小さいと易しい項目ということになる。

a は識別力を表す。識別力とは，項目困難度パラメタ b 付近の曲線の勾配の大きさに関係する指標である。左の項目の方が，勾配が急である。すなわち，その前後で正答確率の変化が大きい。言い換えれば，その値付近では，特性尺度値の微小な違いが正答確率に極めて敏感に影響することを意味する。

1 パラメタ・ロジスティック・モデルでは (2)式のように項目パラメタの数が 1 個減る。項目特性関数の a に項目の違いを表す添え字「j」がないことが違いである。すなわち，全ての項目で a パラメタの値が共通であることを意味している。

$$P_j(\theta) = \frac{1}{1 + exp\{-1.7a(\theta - b_j)\}} \tag{2}$$

1 パラメタ・モデルの当てはまりが良いのは，古典的テスト理論の項目分析で識別力を表す点双列相関係数の値が項目間であまり違わない場合である。実際にモデルを適用する場合には，現実とのバランスを探っていく必要がある。いくらでも複雑なモデルを作って現実のデータへの当てはまりを良くすることはできるが，それだけでは意味がない。現実に結果として有用な結論が出せるかどうかが重要と言える。現実問題として，受験者数が少ない場合には，パラメタの数の少ない方が，推定精度が良くなるので，1 パラメタ・ロジスティック・モデルの方が有効に機能する場合がある。

3 パラメタ・ロジスティック・モデルは (3) 式のように，項目パラメタ数が 3 個となる。2 パラメタ・モデルと同じ a_j，b_j にパラメタ c_j が加わる。c_j は項目特性曲線の下方漸近線の値を示す。3 パラメタ・モデルの場合は $b_j = \theta$ において，$P_j(\theta) \geq 0.5$ となる。多枝選択形式の問題の場合，あて推量（random guessing）が起きるので，本来は正答できない場合でも偶然に正答する確率が生じてくるのである。その分が上乗せされたモデルである。

$$P_j(\theta) = c_j + (1 - c_j)\frac{1}{1 + exp\{-1.7a_j(\theta - b_j)\}} \tag{3}$$

3 パラメタ・モデルの場合，c_j の推定場面で問題が生じることがある。推定計算の収束が悪い，c_j の推定値が現実に照らして考えられない値になる，といったことがある。さらに，モデル上あて推量が必ず一定の確率 c_j で生起するとしている点も問題である。例えば，特性尺度値が極めて低い受験者が 5 枝選択の問題に解答する場合，あて推量で正答する確率は 0.2 となるかもしれない。全く手掛かりがなければランダムに正答を選ぶしか方法がないからである。しかし，より特性尺度値が高い受験者の場合，完全に正答が分からない状態ではない。例えば，選択枝を 2 つに絞ったとすれば，あて推量による正答確率は 0.2 ではない。それにもかかわらず，3 パラメタ・モデルの場合，θ の全域で常に一定のあて推量による正答確率を与えてしまうのである。

実際のテストでは，3 パラメタ・モデルに矛盾するケースが見かけられる。すなわち，非常に特性尺度値が低い受験者は完全にあて推量で解答するが，特性尺度値が少し高い受験者の正答確率の方が低くなることがある。つまり，受験者が正答を考える状況になると正答確率が下がるが，もう少し理解が深まると，正答確率が再び上がり出すというような項目も稀ではないのである。

そもそも，正答確率がそこまで低くなるような項目をテストに含めること自体に問題があると言える。測定対象に適合した項目とは言えない。現実に適用するには，無理に c パラメタを導入するよりも，2 パラメタ・ロジスティック・モデルの方が実用的と考えられる所以である。

4．4．特性尺度値の推定

前項までは基本的に単一の項目に関する議論を行ってきた。しかし，テストは複数の項目に対する応答を基に，受験者の成績を何らかの形で示すことを目的として実施するものである。IRT の場合，特性尺度値で個人の能力を表すが，実際に観測された項目応答パタンが得られる確率が最大となる θ の値をもって，個人 i の特性尺度値の推定値とするのである。なお，以下の例示では，誤答は「0」正答は，「1」と表している。

例えば，4 項目から構成されるテストがあるとする。これら 4 項目の項目パラメタ値は，a_j は共通で 1.0，b_j は順に-1.5, -0.5, +0.5, +1.5 とする。応答パタンを「1，0，0，0」，「1，1，0，0」，「1，1，1，0」･･･，といった形で表すと 2 値の 4 項目なので，理論的に$2^4 = 16$通りの項目応答パタンが生じる。表 1 には典型的なパタンのみを示した。

表 1. ある 4 項目のテストにおける応答パタンの生起確率

特性尺度値	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0.0
パタン#1(1,0,0,0)	0.071	0.148	0.273	0.407	**0.449**	0.346	0.180
パタン#2(1,1,0,0)	0.001	0.005	0.021	0.074	0.192	0.346	**0.422**
パタン#3(1,1,1,0)	0.000	0.000	0.000	0.002	0.015	0.063	0.180

特性尺度値	0.5	1.0	1.5	2.0	2.5	3.0
パタン#1(1,0,0,0)	0.063	0.015	0.002	0.000	0.000	0.000
パタン#2(1,1,0,0)	0.346	0.192	0.074	0.021	0.005	0.001
パタン#3(1,1,1,0)	0.346	**0.449**	0.407	0.273	0.148	0.071

θ の条件付き正答確率を $P_j(\theta)$ ，誤答確率を，$Q_j(\theta) = 1 - P_j(\theta)$ とする。テストを実施する以前の状態では，θ は未知である。項目数を n とすると，特定の応答パタンを得る確率と θ との関数は以下の（4）式で表される。ここで項目応答理論の重要な仮定である「局所独立の仮定」が導入されている。これは，受験者のある項目に対する解答（回答）は，他のいずれの項目に対する解答(回答）とも独立に生ずる，ということを仮定するが，言い換えると，特性尺度値を固定した時に，ある項目応答パタンが生じる確率は，各項目に対する応答が生じる確率の積で表わされる，ということである。

表 1 に(4) 式で得られた結果を示している。ただし，表 1 では θ を 0.5 刻みで表しているが，実際には θ の連続量である。

$$Prob(\boldsymbol{u_i}|\theta) = \prod_{j=1}^{n} P_j(\theta)^{u_{ij}} \cdot Q_j(\theta)^{1-u_{ij}} \quad (4)$$

例えば, θ が 0.0 のときにパタン#1 の生起確率は 0.180, パタン#2 は 0.422, パタン#3 は 0.018 となっている。縦方向に 16 パタンの生起確率の全てを足すと，1.0 となる。

一方, 実際にテスト結果として得られるのは項目応答パタンである。例えば, パタン#2 では結果は「正答」,「正答」,「誤答」,「誤答」なので，各 10 点の配点の正答数得点で表せば，20 点の得点となる。これに対して IRT の特性尺度値 θ を推定する場合には, この応答パタンが生じる確率が最も高い特性尺度値の値を当該受験者の推定尺度値 $\hat{\theta}$ とする。

すなわち，(5)式を用いて推定する。(5)式は (4)式と右辺が等しいが，(4)式では θ が与えられている時に $\boldsymbol{u_i}$ が得られる確率を表わすのに対して，(5)式では逆に $\boldsymbol{u_i}$ が与えられている時に θ が得られる確率を表わしている。θ が変数であり，(5)式は尤度関数と呼ばれる。

$$L(\theta|\boldsymbol{u_i}) = \prod_{j=1}^{n} P_j(\theta)^{u_{ij}} \cdot Q_j(\theta)^{1-u_{ij}} \quad (5)$$

例えば，θ= -3.0 でこのパタンが生じる確率は 0.001，　θ= -2.5 では，0.005 …となる。そして，　θ= 0.0 で最大値 0.422 を取る。このように，そのデータが生ずる可能性が最も高い尺度値を θ の推定値とするのである。例えば，パタン＃3 であれば, θ= 1.0 で最大値 0.449, パタン＃1 では θ= -1.0 で最大値 0.449 を取る。最大の尤度 を与える θ の値が最尤推定値である。

正答数得点の考え方と比べ, IRT では項目応答パタンをそのまま生かしているという点が利点である。例えば，「1，0，1，0」というパタンも正答数が 2 なので，正答数得点ではパタン＃2 と同じ結果となる。IRT ではパタンが異なると θ の推定値も異なる。そういう意味で, テスト結果が持つ情報をより活かしていることになる。

4．5．項目パラメタの推定

前項では項目パラメタは既知のものとして扱っていたが, 実際のテストでは項目パラメタ値を先ず推定しなければならない。受験者の項目応答行列を基に, そのようなデータが得られる確率が最も高くなるような項目パラメタの値を

計算して推定値とする。同時最尤推定法，周辺最尤推定法，また別にベイズ推定法などがある。項目パラメタ推定法およびそのための数値計算法に関しては最近発展が著しいがここでは省略する。

(6) 式は同時最尤推定法の尤度関数である。$\boldsymbol{\theta}$ と $\boldsymbol{a}$ と $\boldsymbol{b}$ を同時に推定する。実際に必要なのは項目パラメタ推定値であるが，受験者の特性尺度値 θ についても同時に推定するのがこの方法である。ただし，ここで $\boldsymbol{a}$ は識別力パラメタベクトル，$\boldsymbol{b}$ は困難度パラメタベクトルを表わす。

$$L(\boldsymbol{u}|\boldsymbol{a},\boldsymbol{b},\boldsymbol{\theta}) = \prod_{\mathrm{i}=1}^{\mathrm{N}} \prod_{j=1}^{n} P_j(\theta)^{u_{ij}} \cdot Q_j(\theta)^{1-u_{ij}} \tag{6}$$

それに対して（7）式は周辺最尤推定法の尤度関数である。周辺最尤推定法では，受験者の母集団分布を仮定することによって受験者個人の $\boldsymbol{\theta}$ を推定しない。「標本数（受験者数）が多くなると推定値が真値に一致する」という一致性を満たし，同時最尤推定法よりも統計数理的に良い性質を持っている。

$$P(\boldsymbol{u}_p) = \int_{-\infty}^{+\infty} g(\theta) \prod_{j=1}^{n} P_j(\theta)^{u_{ij}} \cdot Q_j(\theta)^{1-u_{ij}}$$

$$L(N_1, N_2, \ldots, N_m|\boldsymbol{a},\boldsymbol{b}) = \frac{\mathrm{N}!}{\prod_{\mathrm{p}=1}^{\mathrm{m}} \mathrm{N_p}!} \prod_{p=1}^{m} P(u_p)^{N_p} \tag{7}$$

ただし，$m = 2^n$ で、p は項目応答パタンで $\mathrm{N_p}$ はその項目応答パタンを示す人数を表す

4．6．テストの測定精度

IRT ではテストの推定精度はテスト情報量で表す。テスト情報量は θ の関数として表され，テスト情報関数（Test Information Function）と呼ばれている。テスト情報関数は（8）式で表される。

$$I(\theta) = D^2 \sum_{j=1}^{n} a_j^2 P_j(\theta)\{1 - P_j(\theta)\} \tag{8}$$

古典的テスト理論の下でテストの精度，品質を表わす指標は信頼性係数という受験者集団に依存した一つの値であったのに対して，IRT では(8) 式で $\mathrm{I}(\theta)$ が θ の関数として定義されるため，特性尺度値によってテストの精度に違いがあることを表現することが可能になる。

図 5 は後述する「日本語 Can-do-statements」という 60 項目から成るテストのテスト情報曲線である。この例で言えば，θ=0.0 付近の受験者に対する

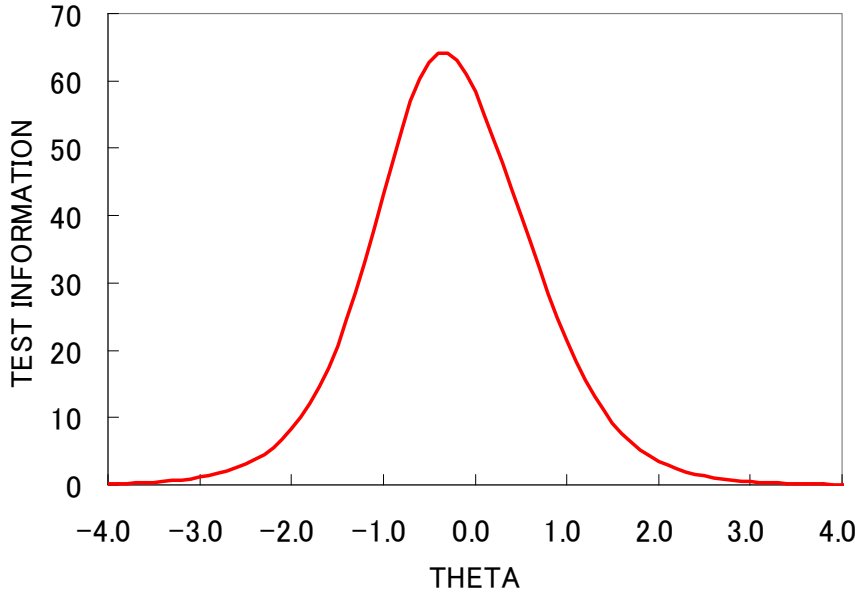

図 5. 日本語 Can-do-statements 60 項目のテスト情報量

特性尺度値の推定精度が極めて高いことが示されている。一方，分布の端は情報量が少なく，推定精度が低い。

例えば，毎年約 50 万名の受験者がいる大学入試センター試験をテスト情報量の考え方で評価するならば，精度が高い「推定」が可能な範囲にはおのずから限界があるということが言える。学力の幅が極めて大きい約 50 万名もの受験者に対して，一種類の試験問題で精度の良い測定が可能なのかどうかという疑問が生じるのである。

4. 7. 等化

複数の IRT 尺度が存在する時，間隔尺度の原点と単位とを揃えた共通尺度を構成する必要がある。 IRT 尺度では，目盛の原点の単位は線形変換の範囲内で自由に設定できる。例えば，温度の単位でファレンハイト(℉: Fahrenheit)とセンチグレード (℃: centigrade)は，原点も単位も異なるが，線形変換を行えば共通の尺度に合わせられる。複数の尺度を共通尺度に合わせることを等化(equating) という。これが IRT 尺度の最大の利点と言える。例えば，テストの実施機会ごとにパラメタを推定し，原点と単位を実施機会ごとに決めていたとしても，適切に線形変換をすることで共通尺度上に乗せられる。例えば，TOEFL-PBT を年間 8 回実施して，各回でパラメタ推定を行って独自の原点

の単位を持つ IRT 尺度を構成していたとしても，複数の線形変換を行うことで，全てを共通尺度上に乗せることが可能となる。

尺度 θ と尺度 θ^*，2 つの尺度があるとする。この 2 つの間に$\theta^* = k\theta + l(k > 0)$という関係が成り立つとすれば，$a^* = a/k$，$b^* = kb + l$ が成立し，結果的に$P^*(\theta) = P(\theta)$ となる。ただし，実際には，k とlをどのようにして推定するかが問題となる。

具体的な等化のデザインおよび k , l の推定に関しては，様々な方法が提案されている。

4．8．特異項目機能

IRT を用いて特異項目機能（DIF: Differential Item Functioning）を検出することも可能である。差異項目機能という訳が用いられることもある。特性尺度値が等しい受験者であっても，属する下位集団が異なると正答確率が異なるという現象が特定の項目で生じている場合，DIF が存在すると言う。

米国社会では，人種や性別で DIF が問題となる場合がある。例えば，受験者が白人でも黒人でも，特性尺度値が等しければある項目に正答する確率も等しくなければならない。もし，白人受験者集団の方が黒人受験者集団よりも正答確率が高くなるようなことがあれば，社会問題化する。意図的ではなくともバイアスを含んだ項目ということになる。

DIF は歴史的には項目バイアスという考え方から始まったが，国際比較をする場合の質問項目の意味的等価性の検討にも用いられている。バック・トランスレーションにより言語的等価性が確認されていたとして（再翻訳法），さらに社会文化的な文脈の中での項目の意味が同じなのか，検討する場合にも用いられる。バイアスという偏ったニュアンスだけではなく，集団によって異なるということを表わす概念であり，価値的に問題のある場合もそうではない場合もある。そのため最近では「項目バイアス」と呼ばれることはなく，一般に「DIF」が用いられている。

図 6 は均一 DIF が生じている項目の例である。特定の焦点集団（Focal Group）が参照集団（Reference Group）に対して，同じ正答確率を得る特性尺度値が一定の値だけずれているような状況である。

図 7 は不均一 DIF が生じている項目の例である。集団による差異が θによって変わっている。図 7 の例では，特性尺度値が低いところでは焦点集団が有利だが，特性尺度値が高いところでは参照集団が有利となって，逆転が起こっ

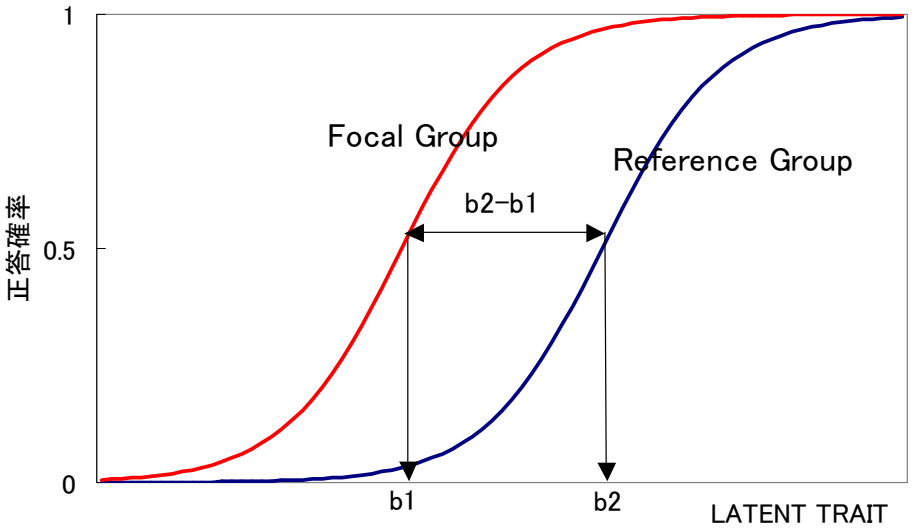

図 6. 均一 DIF が生じている項目特性曲線

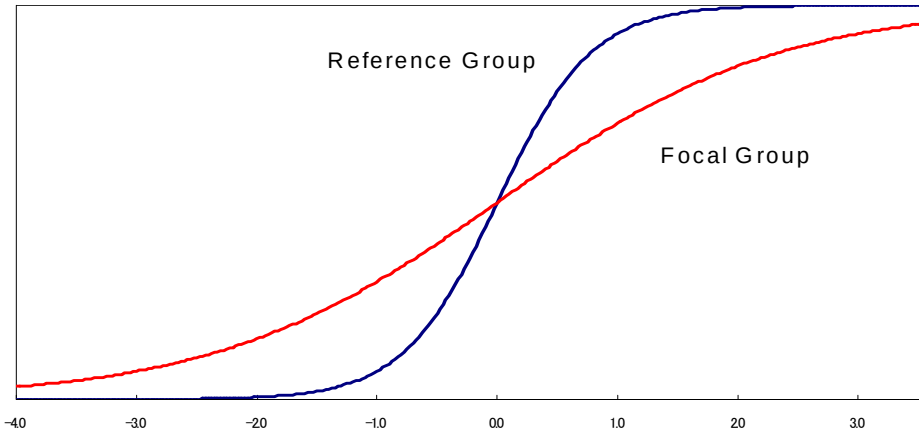

図 7 不均一 DIF が生じている項目特性曲線

ている。

DIF の検出法には大別して IRT を用いるパラメトリック法と特に IRT を用いることのないノンパラメトリック法がある。パラメトリック法ではロードの χ^2 法, 尤度比検定法などがあり, ノンパラメトリック法には Mantel-Haenszel 法, SIBTEST, ロジスティック回帰などがある。

4．9．項目応答理論の限界

IRT は強力なテスト理論であって，実用的にも優れた特徴を持つ。しかし，全てのテストに適用可能というわけではない。精緻なモデルである以上，精緻な仮定が置かれている。したがって，その仮定が満たされない状況で適用しても何も得られないことには留意すべきである。例えば，受験者数が少ないテストやモデルの仮定を逸脱した構造を持つテストに適用することはできない。古典的テスト理論の範囲で十分に有用な情報が得られる場合もある。IRT を適用していない試験が直ちに時代遅れの試験になるということではない。濫用に陥るとむしろ試験の質を低くしてしまう。利点を活用できなければ無意味なのである。

IRT を活用するには，事前に項目パラメタが推定されていることが必要である。すなわち，本番のテストの前に事前の予備テストが必要だが，その際に問題が漏れてしまってはならない。また，良質な項目は再利用される。項目が公開される状況では再利用が不可能となり，IRT を用いた共通尺度化を行うことができない。

さらに，「IRT は絶対評価をするテスト理論である」という受け取られ方をされる場合もあるがそれは事実に反する。共通尺度に乗せられ，共通尺度上の変化が観察可能というメリットはある。しかし，「原点 0 は学力がない」というような絶対評価ではない。原点と単位は任意に定めることが可能だからである。

5．日本語能力の測定に関する諸問題

5．1．項目応答理論を用いた日本語 Can-do-statements の DIF 分析

日本語 Can-do-statements（三枝，2004） とは，1997 年から開発された，日本語学習者の言語能力を測定する尺度である。具体的な場面で日本語を使って行う行動がどの程度できるかを自己評価により測定する。「読む」，「書く」，「聞く」，「話す」の 4 技能に各 15 項目の質問があり，合計 60 項目で構成されている。日本語 Can-do-statements は日本語能力試験の妥当性検討のための外的基準の一つとして用いられることを意図して開発された（島田・三枝・野口，2006）。特定レベルの受験者の日本語運用力について具体的な解釈規準となることをめざしたものである。

この日本語 Can-do-statements に対して，野口・熊谷・脇田・和田（2007）では，IRT 尺度を構成し，母語をもとに下位集団を構成して DIF 分析を実施した。調査対象は，2002 年に国内の日本語学校，大学計 9 校で 794 名の日本

語を学ぶ外国人学生であった。日本語 Can-do-statements で特性尺度値が等しい日本語学習者でも，母語が異なると「できる」と回答する確率が異なる場合，当該項目に DIF が存在することになるが，そのような項目を検出し共通の特徴を抽出することを試みたものである。野口ほか (2007) では母語として中国語と韓国語を取り上げて比較した。日本への留学生はこの 2 ヶ国が多数を占めているからである。

分析の結果，60 項目中 10 項目に DIF が生じていた。中国語母語話者に有利な項目は，全てが「読む」と「書く」技能の項目であった。例えば，「図書館の本棚にある本の背表紙を見て必要な本を探すことができますか」という項目がそれに当たる。漢字の存在が有利に働くとみられる。

一方，韓国語母語話者に有利な項目は，主に「話す」,「聞く」の技能の項目であったが，一部に「書く」の項目も見られた。漢語の存在，発音や専門用語の類似性といった言語的な要因, テレビドラマなどの大衆文化の類似性といった文化的な要因等が考えられる。特に，韓国語の文法構造が日本語に近いことが韓国語母語話者に有利な DIF の出現に影響している可能性がある。

DIF 分析の結果，大部分の質問項目では DIF は検出されなかった。検出された場合には，主として「読む」,「書く」の技能では中国語母語話者が有利に，「話す」,「聞く」の技能では韓国語母語話者が有利な DIF 項目となる傾向があった。

5. 2. パフォーマンステストにおける真正性とその方法論的問題

近年, パフォーマンスを測定するということが重視されている。「言語能力」の場合も実際に使える必要があるということで，実際に「話す」「書く」形式のテストの開発が盛んに試みられている。実際の場面に近い状況での測定ということで, 真正性（authenticity）という概念が強調される（例えば, Bachman, 1990)。単に獲得した知識を問うのではなく，実際の生活場面で他者とコミュニケーションができる，という意味である。従来の客観式のテストは，いわゆる問題項目の個別要素的なテスト（discrete point test)，分析的なテストと呼ばれ，文法・語彙など個々の知識について獲得した程度を重視した測定として批判されることが多くなった。

しかし，パフォーマンスの測定には大きな問題が潜んでいる。まず，全てを実際場面に近づけて行くと，情報が多すぎて「測定する」という視点で大切なものが何か分からなくなってしまう。言い換えると測定したい部分が強調できない。真正性もあくまで言語能力を測定するために必要な概念である。実際の

言語使用場面をそのまま忠実に試験に持ち込むことが適切かどうか，測定目的に照らしてその程度を十分に考える必要がある。あくまでも程度問題として考えるべきなのである。

パフォーマンスの測定には，採点者，評価者の主観の影響が大きい。客観性，信頼性をどうやって確保するのかが大きな問題となる。また，受験者に対して提示できる課題を多くはできず，そういう意味で妥当性の面でも十分ではないこともある。試験の目的によっては完全なパフォーマンステストが適切な場合もあるかもしれない。しかし，測定目的によってはパフォーマンス的な部分と個別要素的な部分とを組み合わせて，信頼性の高い，かつ，言語知識に偏らない試験にすることが適切な場合もある。受験者に学習診断的なフィードバックをしようとするならば，個別要素的な問題項目が必要である。

例えば，ビジネス場面に特化した日本語能力テストで実際のビジネスの場面でのコミュニケーションを評価する場合，ビジネスの典型的場面を含めばそれでよく，こと細かな部分までテスト場面に取り入れる必要はないのである。テストは現実世界の精密な模型である必要はないということである。

パフォーマンステストが抱えた問題の存在は言語テストに限らない。教育の問題では振り子が完全に振り切れるまで新しい考え方を推し進め，その限界を顧みることがないことが稀ではない。新しく出てきた考え方を徹底的に推進し，それまでの考え方を全否定する傾向がある。そういった「振り子現象」が生じやすい。言語教育における「コミュニカティブなアプローチ」が一時期，言語教育，外国語教育，日本語教育などの分野を席巻したが，それが全てであるかどうかを再検討すべき時期に来ているように思われる。大切なのはバランスである。

5．3．パフォーマンステストにおける包括的評価と分析的評価

パフォーマンステストにおいては採点者が主観的に評価を下す必要がある。その場合，包括的評価と分析的評価の 2 つの考え方がある。

包括的評価は，パフォーマンスを要素や側面に分けることなく総体として評価する方法である。それに対し，分析的評価では要素を側面に分けて個別に評価する。例えば，評定尺度を複数用意するようケースである。分析的評価の場合，要素への分割の仕方が重要である。

日本語試験の分野では，双方の事例がある。

包括的評価を採用している試験として，日本語 OPI（Oral Proficiency Interview）がある。OPI とは，外国語学習者の会話のタスク達成能力を，一

般的な能力基準を参照しながら対面のインタビュー方式で判定するテスト(牧野他，2001）である。具体的には，資格を持つ訓練されたテスターが受験者と一対一で面接による直接対話で，質問と応答，ロールプレイなどにより受験者の発話を引き出し，その受験者の話す能力を「初級－下」から「上級－上」「超級」までの 10 段階のいずれかに判定する形式の試験である。判定基準はテスター養成段階で訓練されるが，評定尺度のような形で明示されていない。

一方，庄司・野口・金澤ほか（2004）の「日本語口頭能力試験」は分析的評価を用いている。日本語口頭能力試験は大規模試験の中で実施することを意図しているという制約がある。また，日本語能力試験を構成する「話す」試験を念頭において研究開発が進められた。パーソナル・コンピュータで課題が提示され，それに対する受験者の発話が記録され，それに対して訓練された採点者がチェックリスト評定（言及事項の量的評定)，査定基準評定（質的評定）を行う。日本語能力試験 2 級との相関が 0.674（庄司他，2004)，日本語 OPI との相関が 0.64 程度という結果が出ている。

さらに，海外技術者研修協会（AOTS）の口頭能力試験がある。海外からの技術研修生を対象とした口頭能力試験（Shoji, et al., 2004）である。このテストは日本語口頭能力試験に近い仕様であり，分析的評価となっている。

包括的評価ではどのようにして客観性を保つか，分析的評価では，個々の評価をどのようにして総合的に要約するのかが問題となるが，いずれも難しい課題である。

5．4．解釈基準

受験者の成績からその受験者の能力水準を判断するために，解釈基準が必要である。どのような解釈基準が必要であるかは試験の測定目的によって異なるが，現状では解釈基準が存在しないテストが少なくない。

例えば，TOEFL の解釈基準はホーム頁上には公表されておらず，現段階では存在しないと思われる。受験者の得点は，受験者集団の中の相対的な位置として，解釈することになるが，TOEFL は入学試験で用いられ，そのため競争選抜としての側面が強く，受験者集団の相対的位置が重要な情報である。

TOEIC の場合，受験者の個人の得点に加えて，Score Descriptors Table（レベル別評価の一覧表）として，得点範囲に対して，その範囲に入る受験者の長所および弱点が具体的に記述されたものが解釈基準として用意されている。

BJT 日本語ビジネステストの場合も，BJT の評価結果が実際にどのような日本語運用能力を反映したものなのかを明らかにするために，受験者が実際の

場面で出会う言語行動を記述した質問項目に対して日本語でどのくらいできるかを 5 段階で自己評価した結果をまとめて CANDO レポートとして公表している。

欧州言語テスター協議会（The ALTE: The Association of Language Testers in Europe），はテスト開発者までを含んだ学際的な団体であるが，欧州内で使用されている言語に対して共通の枠組（framework）に基づく Can-do statements を作成している。それを扇の要として，欧州各国で実施されている各言語テストに適用し，異なる言語テストで設定されるレベルの互換性を持たせようとしている。複言語主義に基づき，欧州評議会（Council of Europe）に加盟している各国の作成した言語テストによる結果を相互比較できる共通枠組み CEFR（Common European Framework of Reference for Languages）を作成している。

日本語能力試験では，解釈基準は用意されていないものの，日本語 Can-do-statements（三枝他，2004）によって一定の成果が得られている。

5. 5. 日本語能力測定の今後の展開

最後に，外国語としての日本語試験の今後について言及する。日本語に関する試験では，世界各地の国・地域の受験者が想定される。すなわち，試験に対する評価にも国際的な基準が適用されることになる。テスト理論的な面では項目応答理論を活用した尺度が必要となる。言語教育的な面では，試験の仕様にコミニカティブな要素と構造的な側面をバランスよく取り入れていかなければならない。例えば，日本語の文字はヨーロッパ系の言語にない特徴と言える。文字に関連した知識や技能の問い方には独自の配慮が必要である。さらに，解釈基準の内容も国際的評価にさらされる。試験の社会的責任や試験に関する情報の公開性，透明性をできるだけ高めていくことが必要となる。

試験開発に関するモデルは，主に欧米を中心として研究されてきた。例えば，欧州評議会の CEFR を日本語能力試験に適用すれば日本語能力試験の結果を国際的な解釈基準の中で位置づけることが可能となる。その際に文字，それも表意文字と表音文字とが混在する日本語の特徴を十分に踏まえて，どのような調整をするかが大きな課題となるであろう。

日本には独特のテスト文化がある。例えば，非公開の試験問題であっても，何らかの方法で復元されて販売されてしまう。高い質のテストを開発・実施し続けるためには，予備テスト，尺度の等化，問題項目の再利用などが必要であるが，そのためには問題項目を非公開とすることが望ましいが，それが現実に

実効性を持つかどうかは難しい課題である。

今後，国際的評価に耐えられる日本語試験を開発するには，このような困難な諸課題を一つ一つ解決していく必要がある。

参考文献

Bachman, L. F. (1990). Fundamental Considerations in Language Testing, Oxford University Press.

Gulliksen, H. (1950) Theory of Mental Tests, John Wiley & Sons.

Lord, F. M. (1952). A Theory of Test Scores, Psychometric Monograph (No.7), Psychometric Society.

Lord, F. M. & Novick, M. R. (1968). Statistical Theories of Mental Test Scores, Addison-Wesley.

牧野成一・鎌田修・山内博之・斉藤眞理子・荻原稚佳子・伊藤とく美・池崎美代子・中島和子 (2001). ACTFL-OPI 入門, アルク.

McNamara, T. (1996). Measuring Second Language Performance, Longman.

日本語能力試験企画小委員会口頭能力試験調査部会 (2003). 口頭能力試験科目の創設に向けて, 国際交流基金.

野口裕之 (2001). 項目応答理論とその適用 ―日本語能力試験の分析と日米比較調査の DIF 項目検出ー，計測と制御，40-8，555-560，計測自動制御学会.

野口裕之・熊谷龍一・脇田貴文・和田晃子 (2007) 日本語 Can-do-statements における DIF 項目の検出, 日本言語テスト学会研究紀要, 10, 106-118.

大友賢二 (1996). 項目応答理論入門，大修館書店

三枝令子ほか (2004). 日本語 Can-do-statements 尺度の開発，科学研究費補助金研究成果報告書

庄司惠雄・野口裕之・金澤眞智子 ほか (2004) 大規模口頭能力試験における分析的評価の試み 日本語教育, 122, 42-51.

Shoji, Y. Noguchi, H. & Haruhara, K. (2004). Assessing the Potential of A Large-Scale Oral Proficiency Test Using A Checklist, The 12th Princeton Japanese Pedagogy Forum Proceedings, 150-159.

島田めぐみ・三枝令子・野口裕之 (2006) 日本語 Can-do-statements を利用した言語行動記述の試みー日本語能力試験受験者を対象としてー, 『世界の日本語教育』, 国際交流基金, 79-92.

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理想像--』 (Rasch Modeling for Objective Measurement) 関西大学出版

渡辺直登・野口裕之 (1999). 組織心理測定論 ー項目反応理論のフロンティアー，白桃書房