項目応答理論
――TOEFL・TOEIC 等の仕組み――

Principles and Selected Applications of Item Response Theory

大友賢二

Abstract

本稿では，従来の「古典的テスト理論」の問題点を指摘し，最近の「項目応答理論」の基本的考え方を紹介する．内容としては，1.「テスト得点の不思議」として，困難度，得点と能力，logit score の意味，2.「古典的テスト理論」として，基本モデル，信頼性，妥当性，3.「項目応答理論」として，その特徴，モデル，不変性の証明などを取り上げ，最後に，4.「TOEFL・TOEIC 等への活用」として，等化と適応型テスト，CAN-DO statements，分割点の設定など項目応答理論の利用の基本を紹介する．

キーワード：古典的テスト理論，項目応答理論，logit score，等化，適応型テスト，分割点

1. テスト得点の不思議

1.1 テストの困難度

世界で約 130 か国，6,000 以上もの大学等で利用されている TOEFL（Test of English as a Foreign Language）や約 60 か国，年間約 450 万人によって受験されている TOEIC（Test of English for International Communication）を開発・実施しているのは，米国の ETS（Educational Testing Service）である．以下の英文は，その ETS でテスト研究の中心的役割を果たし，2000 年 2 月 5 日に亡くなられた Frederic M. Lord という方の著書の中の 1 節である．

> Proportion of correct answers in a group of examinees is not really a measure of item difficulty. This proportion describes not only the test item but also the group tested. This is a basic objection to conventional item analysis statistics.[1]

テスト項目の困難度はどのようにして求められているのだろうか？100 名の受験者がいて，そのうち 85 名が正解した場合は，その困難度は 0.85 であり，易しい問題と解釈されるのが普通である．もし，35 名しか正解できなかった場合には，難しい問題と解釈される．けれども，テスト項目の困難度は，そのように，ある受験者集団における正答者の割合と考えてよいのだろうか，というのがこの問題提起である．その割合は，テスト項目のことだけではなく，テストされた受験者集団についての情報も表しているのではないかというのである．同じ項目でも，受験者集団が変われば，その困難度は変わってくるからである．この問題提起は，これまで行われてきた項目分析に対する根本的な反論を意味するものである．

1.2 得点と能力

テスト得点と一般にいわれているものは，テスト全体の中の「幾つの項目に正解したか」ということを基盤として求められている．これは，「正答数に基づく得点」（NRS：Number-right Score）と呼ばれているものである[2),3)]．また，「素点」（raw score）とも呼ばれている．しかし，それ以外にテスト得点として提示できるものがあるだろうか？仮に，この NRS が問題なのだといわれても，なぜ問題なのか？という質問に答えることは極めて困難である．

「正答数に基づく得点」は，幾つの項目に正解したかという頻度の合計であって，それ以上の意味は持っていない．そのことに注目しなければならない．これは，美人コンテストの順位を示すこととさほど違いはない．1 位は 2 位より，2 位は 3 位より人気を示す得点が大きいが，1 位と 2 位との差は，2 位と 3 位との差に等しいとは限らない．つまり，3 人の美しさの差は，等しいとい

本会誌では，①文部省（文部科学省）「学術用語集電気工学編」，②本会編「電子情報通信用語辞典」，③本会編「エンサイクロペディア電子情報通信ハンドブック」に基づき用語を統一しており，「Item Response Theory，略称 IRT）」は上記②に従って「項目反応理論」としているが，本稿では著者の希望により「項目応答理論」で掲載した．

大友賢二　筑波大学名誉教授
Kenji OHTOMO, Nonmember (Professor Emeritus, University of Tsukuba, Nishitokyo-shi, 202-0004 Japan).
電子情報通信学会誌 Vol.92 No.12 pp.1008-1012 2009 年 12 月

う保証はどこにもないのである．したがって，得点が85点といっても，それが示す能力の大きさは，正確には不明である．これは，測定尺度のタイプでいえば，「順序尺度」とみなされる．測定尺度は，区別性を持つ「名義尺度」，区別性と順序性を持つ「順序尺度」，区別性，順序性，等間隔性を持つ「間隔尺度」，そして，区別性，順序性，等間隔性，絶対ゼロ点を持つ「比例尺度」があるが，その中の順序尺度と考えられる．

1.3 logit score の意味

学期初めと終わりの２度，同じテストを行った場合，最初のテストで高い成績を取っている受験者は，学期終わりのテストでは，下がる傾向にあるといわれている．これに関する説明の一つとしては，イギリスの統計遺伝学者 Francis Galton（1822-1911）が提案した，集団の平均値に近づこうとする「回帰効果」（regression effect）を挙げることも可能である．

しかし，更に，考えなければならないことは，「正答数に基づく得点」の持っている特質の限界である．例えば，100 点という天井近くで起こる「天井効果」（ceiling effect）と呼ばれているものである．100 点という天井があるために，100 点近辺では，例えば，120 の能力でも 100 と示され，尺度がゆがんでしまうという現象である．この問題解決には，素点を等間隔性の特質を持つ「間隔尺度」に変換することが必要である．

ゆがめられた尺度を生む NRS を正しく解釈するためには，logit score（$ln(r/(MS-r))$）と呼ばれているものが考えられている．これは，“log odds units”を略したものであり，その発音は，文献（４）で示されているように，［LOH-jits］と第１音節に強勢がある．ここでは，ln＝自然対数，r＝素点，MS＝満点を示し，この logit score は，「間隔尺度」の特質を持ち，しかも，天井効果・床面効果のゆがみを避ける力を持つものである．

例えば，表１では，５点しか向上が見られない山田君の NRS を logit score で表すと，0.462 の向上であることが理解できる．佐藤君の 40 点から 50 点までの 10 点の NRS による向上点は，logit score で求めれば，0.405 しかないことが分かる．つまり，NRS で５点しか向上しなかった山田君を，10 点も向上した佐藤君と比べて，努力不足などと判断するのは適切でないことが理解できる．NRS の得点差は，見かけ上のことであることを明らかにしている．文献（５）でも述べているように，このことは，「正答数に基づく得点」に含まれている大きな問題点の一つを指摘しているといえよう．

表１　NRS と logit

佐藤		山田	
NRS	logit	NRS	logit
50	0.000	90	2.197
40	−0.405	85	1.735
差　10	0.405	5	0.462

2.　古典的テスト理論

2.1　基本モデル：観測値＝真値＋誤差

一貫して同一の得点が得られる度合いはテスト得点の「信頼性」と呼ばれている．古典的テスト理論（CTT：Classical Test Theory）の主な関心事は，この信頼性にあったといえる．そして，我々が「テスト得点」（test score）といっているものは一体どんなものかということに関する検討から「古典的テスト理論」の開発は始められた．その基盤となったモデルは以下のものである．

$$X=T+E \qquad (1)$$

X は「テスト得点」（test score），または，「観測値／測定値」（observed score）と呼ばれるものである．T は「真値」（true score），E は「誤差」（error score）と呼ばれているものである．つまり，テスト得点は真値と誤差からできているとする．こうした古典的テスト理論に関する考察は，多岐にわたっているが，主な事項は，文献（６）でも示されているように，平行測定モデル，本質的に τ 等価測定，同族測定，信頼性係数を決める要因（テストの長さ，項目の識別力，特性値の分散）などを挙げることができる．

2.2　テストの信頼性

信頼性というのは，非常に簡単にいえば，テスト得点の安定性を指している．つまり，あるテストにおける個人の測定結果の一貫性であり，信頼性係数（reliability coefficient）でその度合いが示されている．

古典的テスト理論の多くは，その信頼性が基盤となって考えられている．古典的テスト理論の「信頼性」（reliability）と「測定誤差」（error of measurement）の算出は，以下の式で考えられている．「信頼性係数」は，真値の分散に対する観測値の分散の比で示されており，関係するモデルは以下のとおりである．

信頼性係数＝（真値の分散）／（観測値の分散）
観測値の分散＝（真値の分散）＋（測定誤差の分散）

上のように真の得点の分散と観測値の分散とを比較するということは，信頼性の概念を理解するには役に立つが，「真の得点」は，現実的には求めることはできない．そこで，多くの方法が開発されてきているが，その中でも，どのような配点の場合でも適応できる信頼性係数は「クロンバックのアルファ係数」（Cronbach’s Alpha coefficient）と呼ばれるものである．詳しい信頼性係数の

求め方，信頼性係数の解釈，テスト項目数と信頼性係数などに関しては，文献(7)などに示されている．

2.3 テストの妥当性

テストの妥当性検討というのは，ごく簡単にいえば，「測定しようとしていることを，そのテストが本当に測定しているかどうか」を検討することである．その度合いは，しかしながら，どのようにして求めることができるであろうか？これまでは，妥当性を①内容的妥当性，②基準関連妥当性，③構成概念妥当性，④表面的妥当性，などに分類して，その検討を，それぞれ行ってきているのが普通である．

歴史的には，こうした種類の妥当性を区別して取り扱っていた時期が極めて長い．しかし，現在では，このように別々の妥当性ではなく，妥当性を単一の概念として考えようという流れになってきている．そのことは，「Standards for Educational and Psychological Testing」の最新版，文献(8)での説明からも明らかである．つまり，妥当性は，単一の概念であるというわけである．妥当性というのは，すべての蓄積された証拠が，提案されている目的に関するテスト得点の意図された解釈をどの程度支持することができるかという度合のことを指しているということである．

3. 項目応答理論

3.1 項目応答理論の特徴

「項目応答理論」(IRT：Item Response Theory) は，初めは，「潜在特性理論」(Latent Trait Theory) と呼ばれていたものである．この理論の基本的枠組みは，文献(9)で示され，文献(10)で，厳密な理論体系が樹立されたといわれている．また，これとは全く独立して，デンマークの数学者の提案である文献(11)も，現在「ラッシュモデル」として，項目応答理論に組み込まれて扱われている．これまでテスト界の大きな注目を浴びたことの一つは，以下の文献(12)の発言であろう．これは，もう，56年も前のことになる．

> Ability scores are more fundamental because they are test independent whereas observed scores and true scores are test dependent.

つまり，受験者の能力を示す「能力値」(ability scores) というのは，これまで古典的テスト理論で述べられている「観測値」や「真値」とは，異なるものであるということである．観測値や真値は，受験するテストによって変わるものである．求めなければならない得点は，「受験するテストに依存しない能力値」でなければならない，ということである．

表2 CTT と IRT の比較

領域	古典的テスト理論	項目応答理論
テスト得点	順序尺度	間隔尺度
受験者能力	テストに依存	依存しない
項目特性	受験者集団に依存	依存しない
測定の精度	受験者集団全体	受験者個人ごと
等化・CAT	極めて困難	容易に可能

この視点が，項目応答理論の誕生を促すものであったが，古典的テスト理論と項目応答理論を比較してその主な特徴を挙げるとすれば，表2のようになる．

3.2 モデルの様々

古典的テスト理論に用いられている「正答数に基づく得点」には，多くの限界が見られる．例えば，NRS が35点であった場合，その原因は，受験者の能力が低いからなのか，それともテスト問題が非常に難しかったからなのかが分からないということである．したがって，この二つの要素，能力と困難度を切り離して検討できる道の開発が必要であった．

そこで，非常に簡単にいえば，能力を θ，項目困難度を b，正答確率を P とした場合，$P=\theta-b$ の関係が成立することを目指さなければならなかった．能力が項目困難度よりも大きければ，正答確率は高い．逆に，項目困難度が能力より大きければ，正答確率は低くなるという関係を示すモデルが求められたのである．また，先にも述べたように，得点の天井効果や床面効果をなくし，無限大に能力を表現できるように得点を用いるなどの準備が必要であった．

様々な検討の結果，次のようなモデルが開発されたのである．まず，①ロジスティックモデル (logistic model)，②段階応答モデル (graded response model)，③部分採点モデル (partial credit model)，④名義応答モデル (nominal response model)，などである．

ここでは，このうち，ロジスティックモデルのみを取り上げてみる．今，〈P〉を probability：正答確率，〈exp〉は exponent：(累乗の)指数，〈θ〉は theta：能力パラメータ，〈a〉は item discrimination parameter：項目弁別力パラメータ，〈b〉は item difficulty parameter：項目困難度パラメータ，〈c〉は擬似チャンス水準パラメータ：pseudo-chance-level parameter (いわゆる，「当て推量パラメータ」：guessing parameter)，〈D〉は尺度係数：scaling factor = 1.7，を示すものとする．文献(13)でその詳細を見ることができるが，よく使われるモデルは，以下のようなものである．

・1 パラメータロジスティックモデル

$$1\text{PLM}: P=1/(1+exp(-(\theta-b)))$$

・2 パラメータロジスティックモデル

$$2\text{PLM}: P=1/(1+exp(-D*a*(\theta-b)))$$

・3　パラメータロジスティックモデル

$$3PLM: P=c+(1-c)*(1/(1+exp(-D*a*(\theta-b))))$$

3.3　不変性の証明

IRT の特徴の最も重要な事項の一つは，「不変性」(invariance) と呼ばれているものである．ごく簡単にいえば，test-free person measurement，つまりどんな異なったテストを用いても，共通の尺度で能力測定が可能であるということである．更に，sample-free item calibration，つまりどんな受験者集団に実施しても，項目特性に関する共通の値を求めることができるということである．CTT ではこの不変性を保つことは不可能であった．しかし，IRT ではその不変性を保つことがどうして可能なのであろうか？

ここでは，説明が複雑にならないように，最も簡単な 1PLM を用いることとする．今，あるテストを実施したら，次のデータを求めることができたとする．

・能力下位グループ	θ ：	−3.000,	−2.500,	−2.000
	P ：	0.119,	0.182,	0.269
・能力上位グループ	θ ：	2.000,	2.500,	3.000
	P ：	0.953,	0.971,	0.982

IRT の能力が高い受験者集団でも低い受験者集団でも，このテストは一定の困難度パラメータを保つことができるかを検証するために，先に述べた 1PLM のモデルを変換すると，次のようになる．

$$ln(P/(1-P))=1*(\theta-b) \qquad (2)$$

下位グループでも最も能力の低い受験者のデータを使うと，$ln(0.119/(1-0.119))=(-3.000-b)$

$b=-1.000$

上位グループでも最も能力の高い受験者のデータを使うと，$ln(0.982/(1-0.982))=(3.00-b)$

$b=-1.000$

つまり，能力が高い受験者集団でも，能力が低い受験者集団でも，データが IRT のモデルに適合していれば，項目困難度パラメータは −1.000 と変わらず，不変性を保つことは可能である．どんな異なったテストを用いても，共通の尺度上で能力測定が可能であることの証明も，これと同じ方法で可能である．

4.　TOEFL・TOEIC 等への活用

4.1　等化と適応型テスト

同一の学力・能力を異なる受験者集団に，異なるテストで測定する場合にはテスト得点が互いに交換可能になるような方法を考えなければならない．その方法は，テスト得点の「等化」(equating) と呼ばれているものである．CTT では，こうした問題は解決することは極めて困難であるが，IRT では，可能である．その理由は，項目特性値は，受験する受験者集団とは独立に求められるからである．更には，能力特性値も受験するテストとは独立に求めることが可能だからである．したがって，異なったテストを受験しても，項目特性値が分かっているのであれば，受験者の能力は同一の尺度上に位置付けることが可能である．

更に，IRT を用いれば，「コンピュータ適応型テスト」(CAT： Computerized Adaptive Testing) を行うことが可能である．このテストは，コンピュータ画面やヘッドフォンを通してテスト項目を提示し，キーボードやマウスを用いて行った回答を元に，即座に受験者の能力を推定するのである．そして，項目プールから，受験者の能力水準に適した項目を選び，それを提示し，回答させることによって受験者の能力を推定するというものである．

以上の等化の利用は，TOEFL と TOEIC では既に行われている．CAT に関しては，Computer-Based TOEFL の Listening と Structure の部分で利用されている．しかし，現行の Internet-Based TOEFL においては，世界的規模の実施におけるコンピュータ利用技術，応答構築型項目を含む問題形式，等の課題を解決すべく，CAT の利用は目下鋭意準備中の模様である．

4.2　CAN-DO statements

あるテストでこのくらいの得点であったということは，評価の基準の一つにはなる．しかし，その受験者の能力ではいったい何ができるのかということに関しては明らかではない．そこで，このテストでこの得点であれば，こういうことができるであろうということを予測するための判断材料が欲しいわけである．CAN-DO statements は，そのために，Common European Framework of Reference for Languages，日本英語検定協会の英検など，多くのテストで課題として取り上げられている事項である．

TOEFL に関しては，現在行われている TOEFL iBT では，この CAN-DO statements は，competency descriptors という表記で示されている．これは，受験者の自己評価を基盤として作成されている．全体としての能力は，Overall Language Competency Descriptors のそれぞれに関して Score levels (0-120) を，得点ごとに区分し，それぞれの得点を取った場合は，very unlikely，unlikely，borderline，likely，very likely のうちどれに当てはまるかを示している．TOEIC に関しては，文献 (14) などで報告されている．また，最近の研究結果は，文献 (15) にその成果を見ることができる．

4.3 分割点の設定

ある目標が設定された場合，それに到達したか否かを決定することが重要である．この分割点設定（standard setting）に関する研究は，我が国では極めて少ないが，外国での研究に関する進歩の度合いは，目覚ましい．その最もまとまっている最新の出版物は文献(16)であるが，その中で，特に注目を浴びているのは，「しおり推定法」(Bookmark Method）と呼ばれているものである．

この方法の開発者による論文は，文献(17)に見られるが，分割点推定者が，テスト項目冊子に「しおり」を置くことによって分割点を定めるものであり，そこから，「しおり推定法」の名称が生まれたといえる．その特徴は，①ほかの設定法では見られなかった「項目応答理論の利用」を，まず，挙げることができる．②複数の分割点を設定することが可能，③応答選択型項目でも応答構築型項目でも，いずれの場合でも利用することが可能，④設定作業は極めて簡素化することが可能，⑤テスト問題の内容が，分割点設定の作業にも反映できるということである．

テスト項目でどの程度の正答率がある場合にある段階の目標に到達したと判断するのかということは，極めて興味深い課題である．様々な実験結果で今いえることは，0.67 の正答確率を支持する研究が大方である．しかし，その場合，どのような能力の受験者の正答確率を指しているのか？その必要な受験者の能力はどうしたら求めることが可能であろうか？こうした課題の解決にも，項目応答理論は極めて有力な力を与えてくれる．

こうした正答確率を生む場合の受験者の持つべき能力については，文献(18)によれば，次のことがいえる．つまり，1PLMでは，基本のモデルを変換すると，$\langle \theta = ln(P/(1-P)) + b \rangle$となる．したがって，$P$ を 0.67 とすると，$\theta = 0.708 + b$ となる．つまり，受験者が 0.67 の確率で応答選択型項目に正解するために必要な能力は，問題の困難度より 0.708 ロジット大きい値であるということになる．また，2PLM では，基本のモデルを変換すると，$\langle \theta = ln(P/(1-P))/(1.7 * a) + b \rangle$となる．$P=0.67$，$a=0.5$とすると，$\theta = 0.708/(1.7 * 0.5) + b$ となる．したがって，問題の困難度より 0.833 ロジット大きい値が必要な能力となる．3PLM では，基本のモデルを変換すると，$\langle \theta = ln((P/(1-P)) * (1-c) - c)/(1.7 * a) + b \rangle$となる．$P = 0.67$，$a = 2.0$，$c = 0.3$ の場合は，$\theta = ln((0.67/(1-0.67) * (1-0.3) - 0.3)/(1.7 * 2.0) + b$ となる．したがって，問題の困難度より 0.034 ロジット大きい値が必要な能力ということになる．

以上のように必要な能力の値を求めるための基本モデルの変換のうち，1PLM 及び 2PLM に関しては，文献(19)や(20)に示されているが，その変換の詳細な過程が示されていない．しかし，その変換の過程は，項目応答理論理解のためにも，極めて大きな意味を持つものと考えられる．

文　　献

(1) F.M. Lord, Applications of Item Response Theory to Practical Testing Problems, p.35, Lawrence Erlbaum Associates, Publishers, 1980.
(2) F.M. Lord, Applications of Item Response Theory to Practical Testing Problems, p.51, Lawrence Erlbaum Associates, Publishers, 1980.
(3) R.K. Hamleton, H. Swaminathan, and H.J. Roers, Fundamentals of Item Response Theory, p.77, Sage, 1991.
(4) T.F. McNamara, Measuring Second Language Performance, p.165, Longman, 1996.
(5) B.D. Wright and G.N. Masters, Rating Scale Analysis, pp.32-37, MESA Press, 1982.
(6) 池田　央，現代テスト理論，pp.11-27，朝倉書店，1994.
(7) 大友賢二，項目応答理論入門：言語テストデーターの新しい分析法，pp.42-51，大修館書店，1996.
(8) American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME), Standards for Educational and Psychological Testing, p.11, AERA, 1999.
(9) F.M. Lord A Theory of Test Scores, Psychometric Monograph, vol.7, Psychometric Society, 1952.
(10) F.M. Lord and M.R. Nocick, Statistical Theories of Mental Test Scores, Addison-Wesley, 1968.
(11) G. Rasch, Probalistic models for some intellegence and attainment tests, Denmarks Paedogogiske Institute, 1960.
(12) F.M. Lord, “The relation of test score to the trait underlying the test,” Educational and Psychological Measurement, vol.13, pp.517-548, 1953.
(13) 組織心理測定論：項目応答理論のフロンティア，渡辺直登，野口裕之（編著），pp.3-58，白桃書房，1999.
(14) Chauncey Group International Ltd., Can-do Guide: Linking TOEIC Scores to Activities Performed Using English, 1998.
(15) D.E. Powers, H.-J. Kim, and V.Z. Weng, Redesigned TOEIC Test: Relations to Test-Taker Perceptions of Proficiency in English, ETS, 2008.
(16) M.J. Zieky, M. Perie, and S.A. Livingston, Cutscores: A Manual for Setting Standards of Performance on Educational and Occupational Tests, ETS, 2008.
(17) D.M. Lewis, H.C. Mitzel, and D.R. Green, “Standard setting: A bookmark approach,” In IRT-based standard setting procedures utilizing behavioral anchoring, D.R. Green(Chair), Symposium conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ, 1996.
(18) 言語テスト：目標の到達と未到達，vol.2，大友賢二(監修)，中村洋一，小泉利恵（編），p.39, ELPA, 2008.
(19) G.J. Cizek, M.B. Bunch, and H. Koon, “Setting Performance Standards: contemporary methods” In Educational Measurement: Issues and Practice, vol.23, no.4, pp.31-50, 2004.
(20) G.J. Cizek, “Standard Setting,” In Handbook of Test Development, S.M. Downing and T.M. Haladyna, ed., pp.225-258, Lawrence Erlbaum Associates, Publishers, 2006.

（平成 21 年 6 月 30 日受付）

大友　賢二（おおとも　けんじ）

昭 31 東北学院大・文経・英文卒．昭 41 Georgetown 大言語・言語学研究科退学．昭 54～55 UCLA 外国語教育研究科客員研究員．(財)英語教育協議会，神奈川大教授，筑波大教授，常磐大学部長，日本言語テスト学会名誉会長．著書『項目応答理論入門』など．