早稲田大学審査学位論文
博士（人間科学）

潜在ランク理論を用いたコンピュータ適応型テストのためのアルゴリズムの提案と実装

Proposition and Implementation of Algorithm for Computer Adaptive Test Based on Latent Rank Theory

２０１３年１月

早稲田大学大学院　人間科学研究科
木村　哲夫
KIMURA, Tetsuo

研究指導教員：　永岡　慶三　教授

目次

図目次

表目次

序論

・研究の背景

教育にコンピュータを利用する試みは1970年代にすでに始まっているが，2001年に，「我が国は，すべての国民が情報通信技術（IT）を積極的に活用し，その恩恵を最大限に享受できる知識創発型社会の実現に向け，早急に革命的かつ現実的な対応を行わなければならない.」というe-Japan戦略が，高度情報通信ネットワーク社会形成基本法（平成12年法律第144号）に基づいて内閣官房に設置された「高度情報通信ネットワーク社会推進戦略本部（IT戦略本部）」から出されたことを契機に，教育へのコンピュータあるいはIT利用は急速に進展した.

コンピュータあるいはITを利用したテストを学習評価という側面に限って考えると，主に次のような利点と課題がある.

(1) 教員の採点作業軽減：コンピュータでテストを実施できるようになれば，採点作業にかかる時間はほとんどなくなる．ただし，コンピュータでテストを実施できるようにするまでに相当の作業時間を要する．教員・学校間で協同作業ができるような環境を整え，下記の(4)や(5)が実現されることが望まれる.

(2) テスト結果開示の即時性：テスト実施直後にテストの結果を知ることができることは，すぐに学習評価を行い，次の学習・教育に活かせるので，学習者にとっても教育者にとっても，利点は大きい.

(3) 容易なテストデータ収集：項目ごとの応答情報を紙ベースのテストで収集することも不可能ではないが，膨大な時間を要する．コンピュータを利用することで，項目応答データはすぐに入手可能であり，それに基づいて各項目特性を分析し，問題の改良に生かすことができるだけでなく，次の(4)において，それらの情報を蓄積することができる.

(4) テスト項目のアイテムバンク化：単に項目と正答情報だけでなく，困難度などの項目特性を蓄積しアイテムバンク化することで，実施した項目の再利用や共有が可能となる.

(5) テスト項目の共有：一定の基準で分析したテスト項目をアイテムバンク化し共有することができれば，アイテムバンクが充実したものとなり，テスト作成にかかる労力が軽減されるだけでなく，次の(6)も可能となる．ただし，繰り返し使うことの弊害にも十分注意しなければならない.

(6) コンピュータによるアダプティブな出題：どのくらいの項目数が必要かについては意見の分かれるところであるが，ある程度の項目数がアイテムバンクに蓄積されれば，全受験者に同じ問題を解答させるのではなく，各受験者の解答の正誤によって，コンピュータが困難度を調整して出題するコンピュータ適応型テスト（computer adaptive test: CAT）を実施することも可能になる．理論的にCATの方が，全受験者が同じ問題を解答する場合よりも，少ない問題で同程度かそれ以上の測定精度でテスト結果を出せる.

かつては，テストにコンピュータを利用することは相当な費用がかかり，大規模な開発・実施でなければ難しかったが，Moodleを代表とするオープンソースのラーニング・マネジメント・システム（learning management system: LMS）が登場したことと，パーソナル・コンピュータの飛躍的な処理能力向上により，小規模であってもコンピュータを利用したテストの実施が可能な時代となった（Hinkelman & Grose, 2004; 木村, 2008b; Kimura, 2009）．規模にかかわらず誰もがCATを開発実施する時代も，もうすぐそこまで来ていると言ってもよいだろう．

・本研究の目的と意義

本研究の目的は，理論と実践の両面から CAT について検討を加え，新規のアルゴリズムを提案するとともに，オープンソースを利用してそれを実装するシステムを開発し，そのシステムを英語教育へ適用することによって検証することである．本研究では，1 つの教育機関の 1 学年分の学習者を対象とする小規模な CAT 開発を念頭においる．小規模な CAT とは，受験者数が 200 名前後である場合を想定したものである．

より具体的には，次の 3 点について，理論と実践の両面から論じることが本研究の目的である．

（1） テスト理論の変遷と CAT の根源をふりかえり，小規模 CAT 開発に適したテスト理論と CAT アルゴリズムを検討・提案する．

（2） CAT 開発に利用可能なオープンソースにはどのようなものがあるのか整理し，実際にそれを利用して小規模 CAT を実装するシステムを開発する．

（3） 開発したシステムを英語教育への適用することにより検証するとともに，CAT の受験結果に診断的情報を付加する方法を検討する．

これまで大規模な開発と実施でないと CAT を導入することは不可能であると考えられてきたが，本研究によって，オープンソースを利用して小規模であっても CAT を実装するシステムを開発することができることを示すことで，個人あるいは協働する教員・学校間でも CAT が実装される機会が増えることが期待される．また，テスト結果に診断的情報を加えることは，テスト実施者の説明責任を果たす上でも重要な課題である．

・本論文の構成

本論文は，以下「理論編」と「実践編」に分けて構成されている．「理論編」では，まずテスト理論の変遷を振り返り，本研究で使用するテスト理論を絞り込む．さらに，CAT の根源を振り返り，CAT の利点と問題点を整理した上で，CAT のアルゴリズムを整理し，新しいアルゴリズムを提案するとともに，CAT 開発のフレームワークを紹介し，CAT 開発の段階を整理する．

後半の「実践編」では，CAT 開発のフレームワークの各段階でこれまで行ってきた実践的研究を報告し，CAT 開発の実践面での問題について考察を加えるとともに，小規模 CAT 開発の道筋を示す．具体的には，2 値モデルと多値モデルの分析事例，等化によるアイテムバンクの拡充と統合，開発した CAT を実装するシステムを使ったシミュレーションによる CAT 仕様の検討と，

実テストによるアイテムバンクの検証などである．さらに，英語教育のフィールド行われた CAT 受験者に対するアンケート調査から、項目選択ルールについて再考を加え，能力記述文（can-do statement, CDS）による診断的テスト結果の提示と自己評価の問題についても考察を加える．

なお，本論文の一部は，各研究段階の節目で数回に分けて，国内外の学会で発表し，そこで得られた知見を元に，再構成したものである．また，ここで述べる研究成果の一部はすでに学術雑誌に発表している．

I. 理論編

ここでは，CAT 開発に必要とされるテスト理論の変遷を概観し，本研究で使用するテスト理論を絞り込むとともに，CAT の根源を振り返り，CAT の利点と問題点を整理した上で，CAT のアルゴリズムを整理し，新しいアルゴリズムを提案する．

1. テスト理論の変遷

CAT 開発についての考察を始める前に，テスト結果をどのようにまとめるのか，どのように比較するかといったテスト理論について，その理論的変遷を整理する必要がある．その上で，本論文が目指す CAT はどのようなテスト理論に基づくものであるかを明らかにした上で，本論文が目指す CAT 開発の理論的考察に進む.

テスト理論（test theory）とは，「知能，性格，学力など，個人の心理的特性を測定するテストのスコアに関する統計的問題を取り扱う諸理論の総称」（芝ほか編, 1974: 179）であり，「テスト標準化に関する技術体系であり，テストを経年的に運用したり，コンピュータを用いてテストを実行するための必須の方法論である」（荘島, 2009: 23）．テスト標準化（test standardization）は「テストの尺度化（test scaling）とテストの等化（test equating）からなる複合概念であり，CAT 開発に欠くことのできないものである．尺度化はテストの品質を評価し，能力を測定するために必要なモノサシを作るための手続き, 等化はテストの品質を複数のテストの間で統一するために必要な手続きである」（荘島, 2010a: 37）.

初めてテストのスコア（和得点）を科学的対象としてとらえ，次式のようにテストのスコア X を真値 T と測定誤差 E との和として表したモデルに基づく理論は，古典的テスト理論（classical test theory, CTT）と呼ばれる．

$$\boldsymbol{X = T + E} \tag{1}$$

理想的状態でテストが繰り返し実施（回数が増えるにしたがい X の平均は T に近づく）や，平行テストの概念の導入により，テストの信頼性や妥当性の推定法などが研究された．同一集団の受験した異なる試験の等化手法としては，テストの平均（M）と標準偏差（SD）を使って，素点（x_j: j 番目の受験者のスコア）を標準得点（z 得点や T 得点）に変換する方法が示された（式(2)，式(3) 参照）.

$$\boldsymbol{z = \frac{x_j - M}{SD}} \tag{2}$$

$$\boldsymbol{T = 10z + 50} \tag{3}$$

一方，直接測ることができない受験者の能力を潜在変数（latent variable）としてとらえ，ある項目の得点の期待値を潜在変数の関数で表そうとするモデルがある．因子分析（factor analysis: 芝, 1979）や構造方程式モデリング（structural equation modeling: 豊田, 1998）などとともに，潜在変数モデル（latent variable model, LVM）と呼ばれる．どのような関数をあてはめるか，潜在変数の尺度水準をどのようにとらえるかなどによって，様々なテスト理論が存在する．その中で最も代表的なものが，項目反応理論（item response theory, IRT: Lord, 1980; 芝, 1991）である．IRTでは，いくつのパラメータを想定するかによって，モデルが分かれるが，仮定する潜在変数の尺度水準が間隔尺度で，等間隔性を持った連続変数である．その起源とテストによる測定に対するアプローチは異なるが，ラッシュモデル（Rasch model, RM: Rasch, 1960/1980; Bond & Fox, 2007）も，仮定する潜在変数の尺度水準は間隔尺度で，等間隔性を持った連続変数である．これに対して，仮定する潜在変数の尺度水準が順序尺度で，順序性を持った離散変数であるモデルとして，潜在ランク理論（latent rank theory, LRT: Shojima, 2007a）[1]がある．IRT，RM，LRTは，いずれもLVMという統計モデルに属する．本論文ではRMとLRTに基づいたCATの開発と実践的研究を扱うが，対比してその特徴を明確にするために，以下CTT，IRT，RM，LRTについて，その特徴と制約などについて簡単に述べる．

1.1. 古典的テスト理論（CTT）

CTT はスコア（和得点）を出発点とし，(1)式をモデルとし，テスト全体あるいはテスト受験者全体のことを要約したり，比較したりすることを可能にした．最も一般的なテストの要約は *M* と *SD* によって行われ，これらを使った(2)式や(3)式で標準得点化することで等化を行い複数のテスト結果を比較できるようにした．また，CTT は「テストの信頼性についての理論の発展に大きく貢献している」(村木, 2011: 19)．信頼性係数（reliability coefficient）ρは，スコアの分散（σ_X^2）の中に占める真値の分散（σ_T^2）の割合で定義される．

$$\boldsymbol{\rho = \frac{\sigma_T^2}{\sigma_X^2}} \tag{4}$$

しかし，真値は直接測定することができない(1)式で定義された理論上のものなので，σ_T^2はテスト結果から求めることができず，(4)式からρ を計算することはできない．一般的に，ρ は，平行テスト（交換可能な２つのテスト）か，折半法（１つのテストを 2 等分割して 2 つのテストとして扱う手法）により２つのテストの相関係数から推定される．折半法の場合に次式のスピアマンーブラウンの公式が使われる．折半した 2 つのテスト間の相関係数を r とすると，

[1] 2007 年に Neural test theory として The International Meeting of the Psychometric Society で発表された．

$$\boldsymbol{\rho = \frac{2r}{1+r}} \tag{5}$$

により信頼性係数を求める．テスト項目がすべて正誤の２値である場合，クーダー・リチャードソンの公式 20 （Kuder-Richerdson’s coefficient 20: KR20; Kuder & Richardson, 1937），すなわち，

$$\mathbf{KR20} = \frac{\boldsymbol{k}}{\boldsymbol{k-1}}\left(\mathbf{1} - \frac{\sum_{j=1}^{k} \boldsymbol{p_j}\left(\mathbf{1}-\boldsymbol{p_j}\right)}{\boldsymbol{\sigma_X^2}}\right) \tag{6}$$

ただし，k は項目数，p_jは項目 j の通過率，σ_X^2はテスト全体の分散

や，コロンバックのアルファ係数（Cronbach’s coefficient alpha:α; Cronbach, 1951），すなわち，

$$\boldsymbol{\alpha} = \frac{\boldsymbol{k}}{\boldsymbol{k-1}}\left(\mathbf{1} - \frac{\sum_{j=1}^{k} \boldsymbol{\sigma_{Xj}^2}}{\boldsymbol{\sigma_X^2}}\right) \tag{7}$$

ただし，k は項目数，σ_{Xj}^2は項目 j の分散，σ_X^2はテスト全体の分散

がよく使われる．これらによって求められる信頼性係数は，いずれもテスト全体の信頼性・一貫性を示している．

CCT においても，テスト全体ではなく，各項目の特性を分析する方法も示されている．CCT において項目の困難度は，全受験者中の正解者の割合（通過率）で示す．N 人の受験者がテストを受けたとすると，受験者 i が項目 j に対する反応をu_{ij}（正答の場合 1，誤答の場合 0）とすると，項目 j の通過率p_jは，

$$\boldsymbol{p_j = \frac{1}{N}\sum_{i=1}^{N} u_{ij}} \tag{8}$$

と表せる．

項目の識別力のとらえ方は，いくつかあるが，最も一般的に使われるのは，項目得点とテスト全体の得点の相関係数である項目テスト相関（item-total correlation: IT 相関）である．項目 j の IT 相関r_jは，項目 j とテスト全体の共分散をσ_{XjX}，項目 j の標準偏差をσ_{Xj}，テスト全体の標準偏差をσ_Xと表記すると，

$$\boldsymbol{r_j = \frac{\sigma_{XjX}}{\sigma_{Xj}\sigma_X}} \tag{9}$$

により表せる．IT 相関r_jは相関係数なので，[-1.0, 1.0]の区間の値をとる．

IT 相関r_jは項目の識別力というよりも，テスト全体の中における．各項目の適切さを表す指標と考えた方がよいだろう．$r_j < 0$ ということは，その項目に正答した人ほどテスト全体の得点が低く，その項目に誤答した人ほどテスト全体の得点が高くなる傾向があることを示しているので，その項目はそのテストの中で望ましくない（排除すべき）項目であることを示す．IT 相関が正の値でも 0 に近い値であるということは，その項目に正解するかどうかということとテスト全体の得点との間にほとんど関係がないということを意味するので，やはり，その項目はそのテストの中で望ましくないことを示す．一般に $r_j < 0.25$ の場合，少なくともテストを実施した集団に対しては，その項目は何らかの意味で不適切であることが示唆される．その項目をテストから除去ないし改定することを検討すべきである．IT 相関r_jが低い値になるケースは，その項目の困難度がテストを実施した集団に対して不適切（難しすぎてほとんど全員が不正解，あるいは易しすぎてほとんど全員が正解）である場合と，項目自体に何らかの不備な点（テストで測定しようとしている能力とは異なる能力を必要とする可能性があったり，問題としてあいまいなところがあるなど）がある場合に分けられる．前者は問題の不備ではないが，テストを実施する対象者が合っていない状態なので，その項目は異なる能力水準の対象者に実施すべき項目である．後者は問題の改定を必要とする項目である．

困難度（通過率）と識別力を統合した項目分析の方法として，次のような方法でグラフ化することも可能である．１）受験者を上位からいくつかの等しい群に分け，２）項目ごとに各群の通過率を求め，３）項目ごとに横軸に能力群を縦軸に通過率（正答確率）をとり折れ線グラフを書く．これを項目特性曲線（item characteristic curve, ICC）と呼ぶが，後述の IRT における ICC と区別する必要がある．そのため，こちらを設問解答分析図と呼ぶこともある．豊田（2002:7-10）では，50 項目の学力テストの結果を，得点により低得点群から高得点群まで 5 群（0 群～4 群）に等分に分け通過率を求め，ICC を描き，典型的なパタンとその解釈を 7 通り示している（図 1～図 7）．

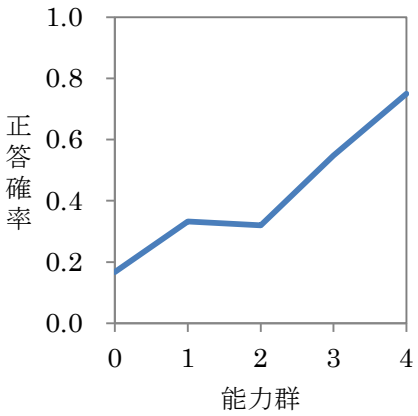

図 1　A 型（模範的な項目の特徴）ICC

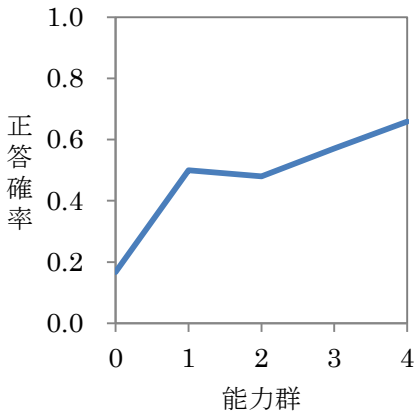

図 2　B 型（低特性者を識別する項目）ICC

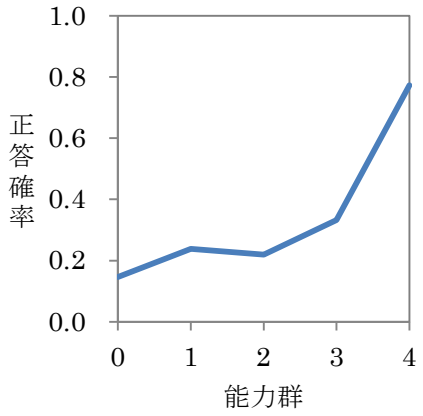

図 3　C 型（高特性者を識別する項目）ICC

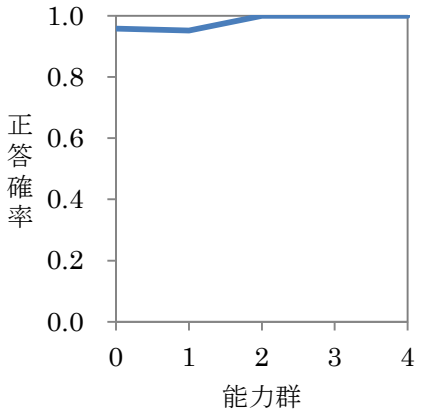

図 4　D 型（易しい項目）ICC

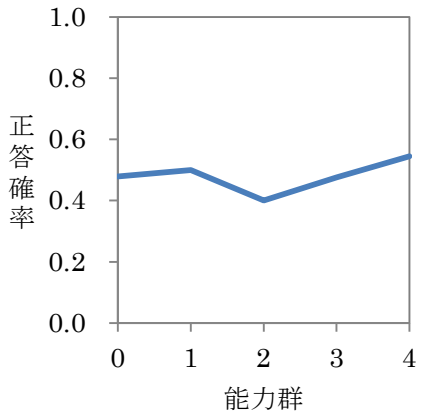

図 5　E 型（中位を漂う項目）ICC

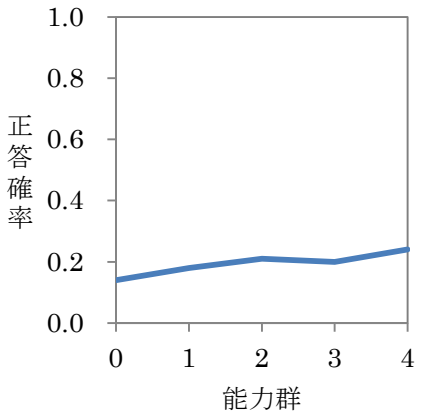

図 6　F 型（難しすぎる項目）ICC

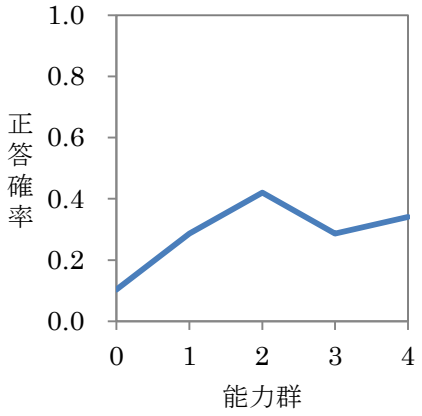

図 7　G 型（右肩下がりの項目）ICC

1.2. CTT の限界

CTT は，*M* と *SD* によるテスト結果の要約，標準得点化による等化，テスト全体の信頼性の分析，正答率や IT 相関による項目特性（困難度と識別力）の数値化など，テストを科学的にとらえる基礎を確立したわけだが，テストの素点を出発点としていることから，テスト理論の目指す標準化と等化において必然的な限界を抱えていた．その限界を打破するために，LVM による IRT や RM が生まれ，さらに LRT が生まれてくるわけだが，ここでは，IRT 等の新しいテスト理論が生まれた背景といえる CTT の限界について整理しておく．

まず，素点（正答数にもとづく和得点）をスコアに使うことに，２つの問題点がある．１つは，スコアの 1 点が測定単位として同一の大きさといえるかという問題である．たとえば，体重計で測定された体重 50kg の人と 55kg の人の体重差 5kg は，体重 80kg の人と 85kg の人の体重差の 5kg と同じであるといえるが，あるテストで素点に基づき測定された 50 点の人と 55 点の人の得点差は，同じテストで 80 点と 85 点を取った人の得点差と同じであるとは言えない．統計学的表現を使うならば，CTT におけるテストスコアは，順序尺度であることは間違いないが，間隔尺度ではないという問題があるということである．CTT では，スコアが間隔尺度だという前提に立って，*M* や *SD* を計算しているわけだが，そのこと自体にすでに問題を秘めているのである．もう 1 つの問題は，CTT におけるスコア 0 点と 100 点の意味である．あるテストで 0 点であるということは，そのテストで測定してようとしている力が全くないということを意味するとは限らない．また，100 点の場合もテストのスコアとしては最高得点であっても，そのテストが測定しようとしている能力の最大値であるといことは意味しない．言い換えると，CTT におけるスコア 0 点と 100 点の意味するところは，そのテストではその受験者の能力を測定できないということである．0 点を取った受験者にとっては，そのテストは難しすぎただけ，100 点を取った受験者にとっては，易しすぎただけということかもしれない．

次に，項目の困難度あるいはテストの困難度についての問題である．CTT において項目困難度は(8)式によって求められることからもわかるように，受験者集団の能力(特性)の分布に依存している．同一のテストでも，優秀な受験者が多く含まれる集団に実施した場合は困難度が低く，優秀な受験者が少ない集団に実施した場合は困難度が高くなる．このことを，「テストの問題の困難度は標本依存である（sample-dependent）」という．CTT においても，(2)式や(3)式によって異なるテストを等化できると説明したが，そこには「同じ母集団から選ばれた標本」が受験した異なるテストであるという前提が必要である．項目の識別力として取り上げた IT 相関も，それを求める(9)式を見ればわかるように，項目の困難度と同様，標本依存である．また，CTT における信頼性を定義した(4)式にも，テストを受けた集団の分散が使われていることから，信頼性も標本依存であるといえる（信頼性を求める公式(6)や(7)を見ても，式の中に標本の分散が使われている）．

第 3 に，受験者の能力の判断はテストの困難度によって左右されてしまうという問題がある．このことを，「能力の決定はテスト依存である（test-dependent）」という．テストが易しい問題ならスコアは高く，難しい問題ならスコアは低くなってしまう．(2)式や(3)式によって，集団の中

で位置づけ(平均からの隔たりやトップからのパーセンタイル)を比較することは可能であるが,そのこととテストが測定しようとしている能力の判断が同じことだとは言えない.

その他にも，測定の標準誤差（standard error of measurement: SEM）も，CTT においては(1)式をモデルとしているので，テスト全体でひとつ（全受験者に同一の）値が示されるだけであることも，CTT が克服できない問題である．信頼性係数についても，同様である．CTT においては，テストの精度を示す SEM も信頼性も,受験者ごとに算出することはできないということである．

以上の CTT の限界を克服するために生まれ，発達してきたのが LVM のもとで生まれた IRT や RM であり，LRT である．

1.3. 項目反応理論（IRT）

IRT は，受験者の能力を潜在変数（latent variable）としてとらえ，ある項目の得点の期待値を潜在変数の関数で表そうとするモデル（LVM）のひとつであり，この点については後述の RM や LRT と共通している.「ある項目の得点の期待値を潜在変数の関数で表す」ということがどういうことか，CTT と決定的に違うのはどこなのかを理解するには，ICC の描かれ方の違いを見るとよく理解できる．CTT では，図 1～図 7 のように，横軸に能力特性群（テストの得点から恣意的に 5 等分に分けたグループ）を置いている．このグラフの描き方は，標本依存であり，テスト依存である．一方，IRT の ICC は，横軸にテストが測定しようとしている受験者の能力の潜在変数を取るので,－∞から＋∞まで理論的にカバーすることになる.一般的に潜在能力を θ とし，θ=0 を横軸の中心におき，－3 から+3 ないし－4 から+4 の範囲で描く．また，横軸は連続する変数なので，グラフは折れ線ではなく曲線で描かれる．縦軸はいずれの場合も正答確率である．

代表的な IRT のモデルの 1 つである 2 パラメータ･ロジスティック･モデル(2-parameter logistic model, 2PLM）を例にとって説明すると，受験者の潜在変数（特性値）をθとし，項目 j の得点の期待値（正答確率）を$P_j(\theta)$とすると，ICC を描く項目特性関数は

$$\boldsymbol{P_j(\theta) = \frac{1}{1 + \exp\{-Da_j(\theta - b_j)\}}}\ , \qquad -\infty < \theta < \infty \tag{10}$$

と表される．D は正規累積モデルに近似させるための尺度因子であり D=1.7 のときにθ全域で最も近似させられることが知られている．しかし，現在ではロジスティックモデルを正規累積モデルに近似させることなく，D が省略されることが多い．a_jは識別力パラメータ（discriminancy parameter）b_jは困難度パラメータ（difficulty parameter）といい，それぞれ項目の識別力と困難度を示す.

2PLM で(10)式にしたがって 5 つの項目について，ICC を描いたものが図 8～図 12 であり，それらを 1 つのグラフ上に重ねたものが図 13 である（木村（2008a）のデータの一部を使用）．図 8 は識別力も困難度も中程度の項目，図 9 は識別力が強く困難度は低い項目，図 10 は識別力がやや弱く困難度はやや高い項目，図 11 は識別力も困難度も低い項目，図 12 は識別力が低く困難

度は高い項目である. 図 8 の項目は中程度の能力を識別するのに適した項目であると言えるのに対し，図 9 の項目は下位の能力を識別するのに非常に優れている項目であることがわかる．図 11 の項目も図 12 の項目も識別力が低いのは同じだが，前者は易しい問題で後者は難しい問題であることがわかる．複数の ICC を重ねた図 13 を見ると，曲線が交差している部分があるので，2PLM の分析では，必ずしも常に能力が高い者ほど，すべての困難度の項目において，正答しやすいわけではない状態になっていることがわかる．この現象は，識別力パラメータをモデルに入れない 1 パラメータ・ロジスティック・モデル（1-parameter logistic model, 1PLM）（上記の(10)式において$a_j = 1$とした場合と定義できる）あるいは，後述の RM では起こらない現象である．

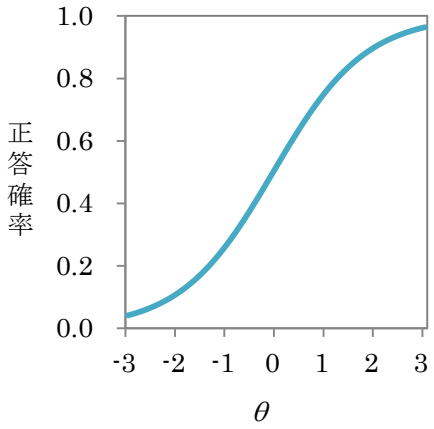

図 8　ICC（a_j=.0.6, b_j=-0.1）

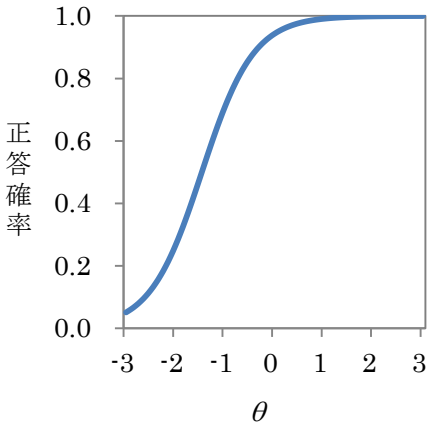

図 9　ICC（a_j=1.1, b_j=-1.5）

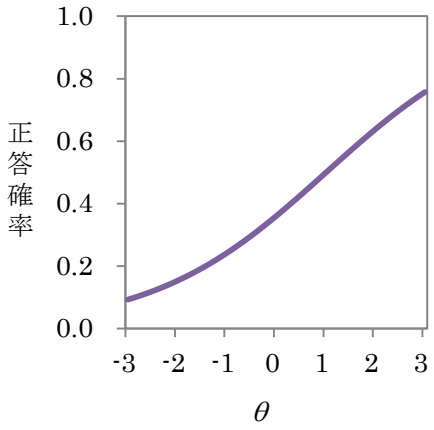

図 10　ICC（a_j=0.3, b_j=1.0）

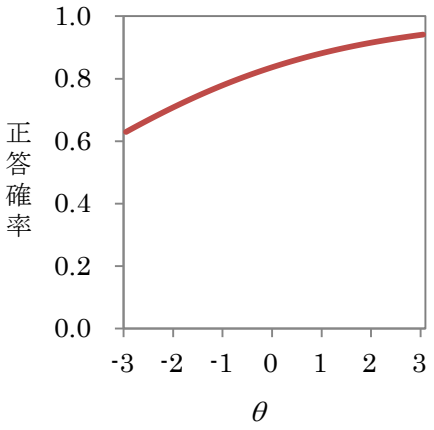

図 11　ICC（a_j=0.2, b_j=-4.4）

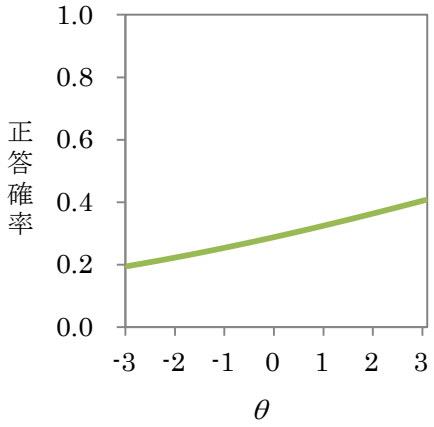

図 12　ICC（a_j=0.1，b_j=5.2）

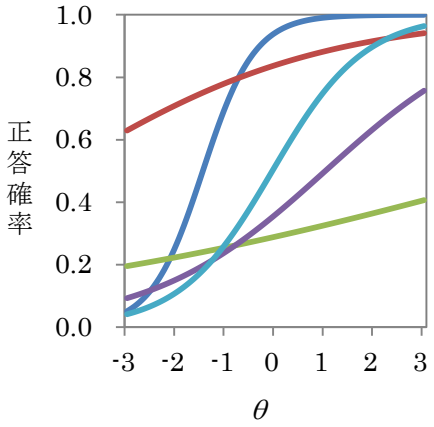

図 13　交差する ICC

IRT はひとつのモデルに集約されているわけではなく，幅広く多様な研究がなされてきた領域なので，その特徴や利点をまとめることは，容易なことではない．大友（1996:17-20）は，CTT の限界との対比で IRT の利点を述べる適切な観点として， Hambleton & Swaminathan (1985:11) の記述を取り上げ，次のように 3 つの利点に焦点を当て整理している．

(1) 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（test-free person measurement）

(2) どんな受験集団に実施しても，共通の項目特性に関する値を求めることが可能（sample-free item calibration）

(3) 測定の精度を能力ごとに算出可能 （multiple reliability estimation）

これらは，どれも前節で整理した CTT の限界の裏返しである．(1)は「能力の決定はテスト依存である（test-dependent）」という CTT の限界を，(2)は「テストの問題の困難度は標本依存である（sample-dependent）」という CTT の限界を，(3)は「テストの精度を示す SE も信頼性も，受験者ごとに算出することはできない」という CTT の限界を IRT が超えたことを意味している． この 3 つの IRT 利点は，後述の RM にも LRT にも当てはまることである．

IRT において，どのようにして能力ごとに測定の精度がわかるかについては，Birnbaum (1968) によって提案された項目情報関数（item information function, IIF）とテスト情報関数（test information function, TIF）について理解する必要がある．ICC の形状からもある程度分かるが，各項目はどの能力水準に対しても同じ測定精度を持っているわけではない．たとえば，図 9 の項目への応答から，θ が-2 から-1 の受験者に対しては測定精度が高いが，θ が 0 以上の受験者に対しては精度が高いとは考えられない. ある項目が θ の尺度上でどの程度の精度を持っているか示すものが IIF であり，項目 j の IIF を$I_j(\theta)$とすると，

$$\boldsymbol{I_j(\theta) = \frac{P'^2_j(\theta)}{P_j(\theta)Q_j(\theta)}} \tag{11}$$

と表される．ただし，$P'_j(\theta)$ は$P_j(\theta)$ の導関数，$Q_j(\theta)$ は誤答確率で，$Q_j(\theta) = 1 - P_j(\theta)$で求められる．2PLM の場合，$P_j(\theta)$ の導関数$P'_j(\theta)$ は$Da_jP_jQ_j$となることがわかっているので，これを(11)式に代入すると，

$$I_j(\theta) = D^2 a_j^2 P_j(\theta) Q_j(\theta) \tag{11'}$$

のように定式化されている．これに基づき図 8～図 12 の項目の IIF を求めてひとつのグラフに図示したものが図 14 である．

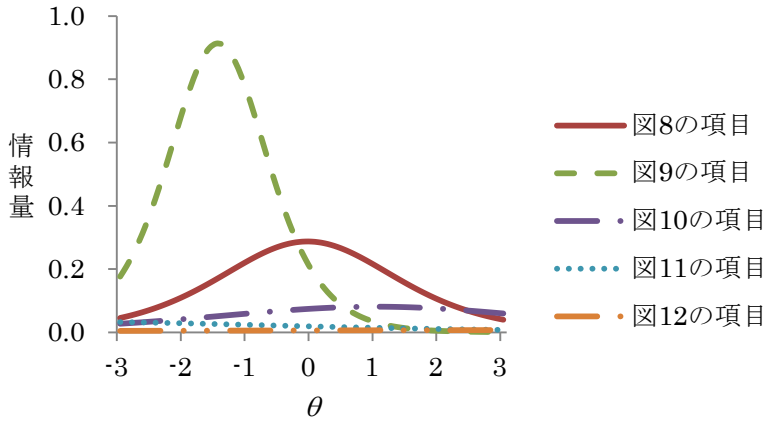

図 14　図 8～図 12 の 5 項目の IIF

式(11')の値はどのような時に大きくなるか（情報量が最大になるか）というと，D^2は定数なので影響がなく，a_j^2は識別力なので，識別力が大きい項目ほど情報量が多くなり，$P_j(\theta)Q_j(\theta)$ は正答確率と負正答確率の積なので，$P_j(\theta) = Q_j(\theta) = 0.5$ の場合が 0.25 に最大になる．つまり，正答確率が 50%に近い項目（b が θ になるべく近い項目）で，なるべく識別力が大きい項目が情報量を最大にする．このことは，後述の CAT の項目選択において利用される情報であり，重要な概念である．

一方，テスト全体が θ の尺度上でどの程度の精度を持っているか示すものが TIF であり，TIF は IIF の単純和で，

$$\boldsymbol{I(\theta) = \sum_{j=1}^{k} I_j(\theta)} \tag{12}$$

と定義される．

1.4. ラッシュモデル（RM）

RM は，1960 年代初頭にこのモデルを発表したデンマークの数学者 G. Rasch の名にちなんでこの名前がつけられている．RM は，項目困難度パラメータbと潜在特性θをロジスティック関数に含むモデルで，

$$P_j(\theta_i) = \frac{1}{1 + \exp\{-(\theta_i - b_j)\}} \tag{13}$$

のように表現される．IRT の 2PLM を表す(10)式のDa_jが 1（識別力がすべての項目において 1）である場合，すなわち 1PLM とみなすこともできるが，その歴史的起源は異なり，測定に対するアプローチが IRT とはまったく異なる．RM に基づいて描かれる ICC は，図 15 のように，すべての曲線が平行になり，2PLM に基づいて描かれた ICC の図 13 のように，曲線がどこかで交差することはない（木村（2008a）のデータの一部を使用）．

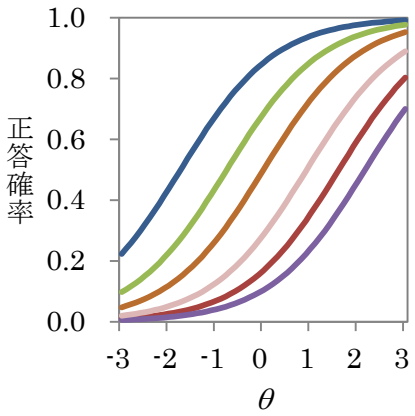

図 15　RM における ICC

RM に識別力パラメータが組み込まれていないのは，モデルを単純にするとか，データ数が少ない場合にも推定ができるようにするといった理由からではない. 識別力などの他のパラメータを組み込むと，客観的測定（objective measurement）を矛盾なく確立することができなくなるからである．客観的測定が確立するということは，テスト中のどの項目においても，潜在能力が高い者が正答する確率の方が，それよりも潜在能力が低い者が正答する確率より，常に大きいということである．このことが RM で成立していて，2PLM では成立していないことは，図 15 と図 13 を比較することでもわかる．

数学的に理解するために，正答確率のオッズがよく使われる．受験者 i が項目 j に正答する確率のオッズを$Odds_j(\theta_i)$とし，別の項目 j' に正答する確率のオッズを$Odds_{j'}(\theta_i)$とすると，これらは，

$$\boldsymbol{Odds_j(\theta_i) = \frac{P_j(\theta_i)}{1 - P_j(\theta_i)}} \tag{14}$$

$$\boldsymbol{Odds_{j'}(\theta_i) = \frac{P_{j'}(\theta_i)}{1 - P_{j'}(\theta_i)}} \tag{14′}$$

と表せる．この 2 つのオッズの比をロジット変換し，RM を表す(13)式を使って展開すると，

$$\boldsymbol{\ln\left(\frac{Odds_{j'}(\theta_i)}{Odds_j(\theta_i)}\right) = \ln\left(\frac{\exp(\theta_i - b_{j'})}{\exp(\theta_i - b_j)}\right) = b_j - b_{j'}} \tag{15}$$

となり，この値は受験者の能力θと独立に常に一定（$b_j - b_{j'}$：2 つの項目の困難度の差）であることが示される．これを，(13)式ではなく 2PLM を表す(10)式を使って展開すると，

$$\boldsymbol{\ln\left(\frac{Odds_{j'}(\theta_i)}{Odds_j(\theta_i)}\right) = \ln\left(\frac{\exp\{-Da_{j'}(\theta - b_{j'})\}}{\exp\{-Da_j(\theta - b_j)\}}\right) = D(a_{j'} - a_j)\theta_i + (a_j b_j - a_{j'} b_{j'})} \tag{16}$$

となり，θを含む項が残るので，この値は受験者の能力θと独立ではなくなってしまう．

同様に，受験者 i が項目 j に正答する確率のオッズを$Odds_j(\theta_i)$とし，受験者 i' が同じ項目 j に正答する確率のオッズを$Odds_j(\theta_{i'})$として，両者の比をロジット変換し，RM を表す(13)式を使って展開すると，

$$\boldsymbol{\ln\left(\frac{Odds_j(\theta_{i'})}{Odds_j(\theta_i)}\right) = \ln\left(\frac{\exp(\theta_{i'} - b_j)}{\exp(\theta_i - b_j)}\right) = \theta_{i'} - \theta_i} \tag{17}$$

となり，この値は項目の困難度bと独立に常に一定（$\theta_{i'} - \theta_i$：2 人の受験者の能力の差）であることが示される．

RM と 2PLM を代表とする IRT のどちらが優れているかについては，簡単に決められない．テストデータへのアプローチの理念に根本的な違いがあり，両者は正反対の理念によって立っている．RM が目指すものは，「客観的な測定結果が得られるようにデータを整理して，意味のある構成概念を作り出すこと」である．これに対して IRT が目指すのは，「手元にあるデータを最大限に忠実に描写するモデルを作り出すこと」である．RM は「モデルにデータを合わせる」ことで客観的な測定結果を得ようとするのに対して，IRT は「データにモデルを合わせる」ことで，そのデータをできる限り説明しようとする（靜, 2007:354）．

ただし，(15)式と(17)式で示された「項目パラメータ推定における標本独立性」と「能力パラメータ推定における項目独立性」は，項目パラメータを推定した時に使われた受験者標本集団の

学力の特徴の影響をまったく受けないということではない. 一度推定されたパラメータ値が不変であり，再推定の必要がないということでもない（村木, 2011: 40-41）.

RM は「モデルにデータを合わせる」ことで客観的な測定結果を得ようとするので，パラメータの推定が終わった後に，モデルの予測値と観測値がどの程度一致するか標準残差（estimated standard residual）z_{ij} が次式により計算され，データのモデルに対するフィットについて検討が加えられる.

$$\boldsymbol{z_{ij} = \frac{u_{ij} - \hat{p}_{ij}}{\sqrt{\hat{p}_{ij}(1 - \hat{p}_{ij})}}} \tag{18}$$

ここで, u_{ij}は受験者 i の項目 j に対して観測された値（2 値の場合, 正解なら 1 不正解なら 0）, $\hat{p}_{ij}$はモデルから予測される正答確率（期待値）を表す. この値は，項目数×受験者の数だけ算出される. 標準誤差（standard error, SE）を要約するために，z_{ij}を 2 乗したものを項目ごとに合計し，受験者数（N）で割ったものが Item Outfit mean square（$Outfit\ MNSQ_j$），すなわち，

$$\boldsymbol{Outfit\ MNSQ_j = \frac{\sum_{i=1}^{N} z_{ij}^2}{N}} \tag{19}$$

であり，z_{ij}を 2 乗したものを受験者ごとに合計し，項目数（k）で割ったものが Person Outfit mean square（$Outfit\ MNSQ_i$），すなわち，

$$\boldsymbol{Outfit\ MNSQ_i = \frac{\sum_{j=1}^{k} z_{ij}^2}{k}} \tag{20}$$

である. この指標$Outfit\ MNSQ$は，z_{ij}の 2 乗の平均を求めたものであり，z_{ij}を求める(18)式の分母は正答確率$\hat{p}_{ij}$が 0.5 のときに最大になり，1 または 0 に近づくほど小さな値になり，正答確率が非常に高い（あるいは低い）ときの意外性により敏感に反応する（$Outfit\ MNSQ$の値が大きくなる）.

$Outfit\ MNSQ$のこの性質を補正するために，考えられたのが Information-weighted mean square（$Infit\ MNSQ$）である. $Outfit\ MNSQ$が重みづけをしないモデル適合指標であるのに対して，$Infit\ MNSQ$は情報で重みづけしたモデル適合指標であり，個々の応答の分散推定値$\hat{p}_{ij}(1 - \hat{p}_{ij})$をもとにした加重平均が計算される. 項目については Item Infit mean square（$Infit\ MNSQ_j$）が,

$$\boldsymbol{Infit\ MNSQ_j = \frac{\sum_{i=1}^{N}(u_{ij} - \hat{p}_{ij})^2}{\sum_{i=1}^{N} \hat{p}_{ij}(1 - \hat{p}_{ij})}} \tag{21}$$

によって求められる．受験者については Person Infit mean square（$Infit\ MNSQ_i$）が，

$$\boldsymbol{Infit\ MNSQ_i = \frac{\sum_{j=1}^{k}\left(u_{ij} - \hat{p}_{ij}\right)^2}{\sum_{j=1}^{k}\hat{p}_{ij}\left(1 - \hat{p}_{ij}\right)}} \tag{22}$$

によって求められる．

いずれの$MNSQ$も，カイ２乗の値を自由度で割ったもので，0 から無限大の値を取りうる．モデルの期待値と観測値が完全に一致している場合 1 となるが，1 より大きい場合は underfit と呼ばれ，モデルから予測できない現象が起きている程度を示す．また，1 より小さい場合は overfit と呼ばれ，測定全体への貢献が少ないことを示す．$MNSQ$の判断基準はいくつかあるが，よく参照されるものとして Bond & Fox(2007)の表 1 がある．

表 1　$\boldsymbol{MNSQ}$の判断基準

Some Reasonable Item Mean Square Ranges for Infit and Outfit (Bond & Fox, 2007:243)

Type of Test	Range
Multiple-choice test (High stakes)	0.8-1.2
Multiple-choice test (Run of the mill)	0.7-1.3
Rating scale (Likert/survey)	0.6-1.4
Clinical observation	0.5-1.7
Judged (where agreement is encouraged)	0.4-1.2

$MNSQ$はモデルの期待値から実際の観測値がどの程度ずれているか，その大きさを示す指標であるが，そのことがどの程度の確率で起こりうるかを示す指標，Outfit standardized fit statistics（Outfit *Zstd*）と Infit standardized fit statistics（Infit *Zstd*）もよく使われる（infit *t* と outfit *t* と呼ばれることもある）．これはMNSQに Wilwon-Hilferty 変換を加え標準化したもので，その値が *t* 分布に従うことが知られている．*Zstd* の値が-1.96～＋1.96（あるいは-2.0～＋2.0）の範囲を外れたものを，ミスフィットとすることが一般的である．ただし，*Zstd* の値はサンプルサイズが大きくなると過剰に反応する傾向があるので，サンプルサイズが 200 を超える場合は注意が必要である．Linacre (2003) は，*Zstd* の値でミスフィットを有効に判断できるのは，サンプルサイズが 100～250 の範囲の場合であることを，図 16 により示している．

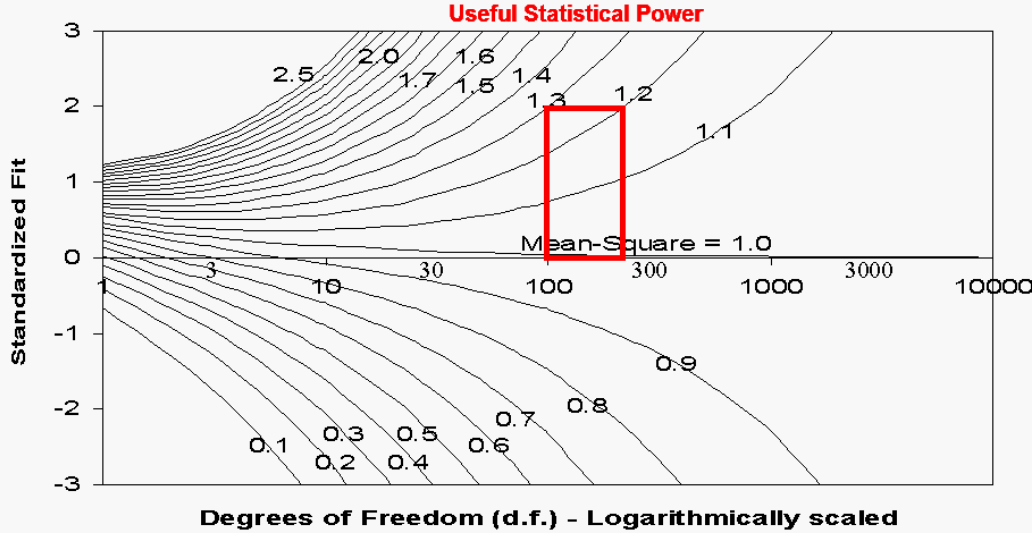

図 16　Size vs. Significance: Standardized Chi-Square Fit Statistic (Linacre, 2003)

*MNSQ*や *Zstd* の値に基づき，モデルに著しく適合しない項目（あるいは受験者）が見つかった場合は，その項目（あるいは受験者）に何かおかしなところがないか，検討を行うべきである．モデルに著しく適合しないと判断した項目（あるいは受験者）のデータを，すべて分析対象から外す方法もあるが，なんらかの基準をもって，明らかに「ケアレスミス」あるいは「まぐれ当たり」と思われる反応だけを取り除く（その部分だけ無回答とする）方法も有効である（Linacre, 2009）．

モデルに適合しないものを除去する場合は，最初の分析で基準に当てはまるものをすべて取り除くのではなく，もっともひどいものから 1 つずつデータセットから除外して再分析を行い，モデルへの適合や信頼性が高まったかをチェックすべきである．ひとつの項目あるいは一人の受験者を除去するだけで，他の項目（あるいは受験者）のモデル適合の指標が大きく変わることがあるからである．不適合なものをどこまで除外するかは，難しい問題であるが，除外して再分析しても改善されなくなった時点で，その除去したものを戻し，ミスフィットの除去を終了させるというのも，現実的で理にかなった方法である（Linacre, 2010）．

1.5.　潜在ランク理論（LRT）

LRTの最大の特徴は，テストの結果をIRTのように連続した細かい値で評価するのではなく，5～10程度の少数の離散的なランクで段階評価するところにある．LRTも，潜在変数を仮定する点においては，IRTやRMと同じであるが，IRTやRMが仮定する潜在変数の尺度水準が間隔尺度で，等間隔性を持った連続変数であるのに対して，LRTが仮定する潜在変数の尺度水準は順序尺度で，順序性を持った離散変数だからである．分析の目的に合わせ，テストの結果をいくつかの段階に分けることで表現する理論であり，いわば段階評価のためのテスト標準化理論である（木村，2010b）．なお，本論文においてLRTは，Shojima（2007a）のニューラルテスト理論（neural test theory,

NTT）のことを指す．LRTは自己組織化マップ（self-organizing map, SOM）や生成トポグラフィックマッピング（generative topographic mapping, GTM）のメカニズムを利用したノンパラメトリック・テスト理論である．SOMによる推定は，ランダムな並べ替えを行っているため毎回の計算が微小に異なるが，GTMによる推定は一括学習型であるため毎回の計算が必ず一致する[2]，推定されるIRP（後述の1.5.1参照）はSOMによる推定の方が少し滑らかである．計算時間はGTMの方がかからないので大規模データに向いている（荘島, 2010b: 98）．本論文で扱うテストデータはサンプルサイズが200前後であることから，SOMによる推定で分析を行った．

LRTにおける離散的なランクのことを，潜在ランク（latent rank）と呼ぶ．「潜在ランク」は，統計的に推定される「学力レベル」「到達度」などと解釈でき，大きいランクに所属している受験者ほど能力が高いことを意味する．テスト結果をいくつの潜在ランクに分析するかは，分析目標とサンプルサイズによって判断される．たとえば，大学に入学してきた学生を，英語の基礎力テストによって3つのレベル別クラスに分けることが目的ならば，潜在ランク数3で分析をする．また，適合度指標や情報量規準（NFI, RMSEA, AICなど10種類）が提案されているので（Shojima, 2008），いくつかのランク数で分析を試み，その値を参照していくつの潜在ランクに分析するのが最もモデルに適合するか判断することもできる．

1.5.1. LRT の項目特性のとらえ方

LRTにおいて，項目の特性は項目参照プロファイル（item reference profile, IRP）で表される．IRPはその項目を受験した場合，各潜在ランクの受験者の正答確率をまとめたもので，グラフ化することで項目の特性を把握しやすい．これはCTTのICC（図 1 ～図7）やIRTの2PLMのICC（図8～図13）と似ているところが多い．特にIRPとCTTのICCは，どちらも折れ線グラフで表現されているので，見た目は区別がつかない．しかし，CTTでは5つの群をテストの総得点で5等分しているのに対して，IRPはSOMやGTMのメカニズムを利用して5つの潜在ランクに分けている点で大きく異なる．

IRTの2PLMにおいて項目の困難度と識別度を表すb パラメータとa パラメータのように，項目の特性を要約するIRP指標も提案されている（熊谷，2007）．IRP指標βとb は項目困難度を表すもので，基準となる値（本研究では0.5としている）に最も近い潜在ランクをβ，その時の値がbである．IRP指標αとa は項目識別度を表すもので，隣り合う2つのIRPの値の差が最大となるペアの若い方の潜在ランクをα，そのときの差がaである．IRP指標γとc は単調増加度を示すもので，隣り合う2つの潜在ランクで正答確率が減少したペア数の割合をγとし，減少した大きさの和がcである．図17中にγ以外のIRP指標を示した（β＝5，b=0.59，α＝3，a=0.23，c=0.10）．γは，隣り合う2つの潜在ランクで正答確率が減少したのは4ペア中1ペアなので，0.25である．

ランク数が上がるにつれて，正答確率も増加するように，IRPの単調増加制約をつけて分析を

[2] 本研究で利用したソフトウェア Exametrika（Shojima, 2010）では，予測できるランダム性を用いてデータの入力順序を制御することで，SOM による推定でも毎回同じ結果が得られるようにされている．

することも可能だが，そうしないことで項目の特性をより柔軟に表現することもできる．たとえば，図17は，ある多肢選択問題のIRPであるが，中程度の潜在ランクの受験者にとって魅力的に思える選択肢があり，どの選択肢も同じに思える下位の潜在ランクの受験者よりも，中程度の潜在ランクの受験者の正答率が低くなっていることを表現している．こういった表現は，CTTのICCでは可能だが，IRTの2PLMやRMのICCでは不可能である．

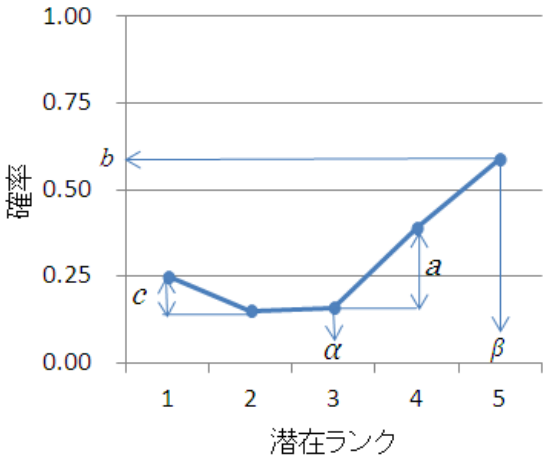

図 17　IRP の例

1.5.2.　LRT の受験者能力のとらえ方

LRTは段階評価なので，テストにより各受験者がどの潜在ランクに属するかを推定する．LRTがIRTと大きく異なるのは，潜在能力を連続変数上の一つの値で推定し，その精度をSEで表現するのではなく，LRTでは，受験者の潜在ランクを順序尺度上に推定すると同時に，受験者が各ランクに所属する確率を集めたランクメンバーシッププロファイル（rank membership profile, RMP）として多義的に表現する点である．これは，他のテスト理論にない，LRTの大きな特徴である．

潜在ランクの推定値が同じだったとしても，RMPの違いによって，受験者に異なるフィードバックを返すことができる．たとえば，図18と図19は同じテストの結果から得られた２つのRMPである．この2つは潜在ランクとしては同じ3だが，図19のRMPの受験者の場合，潜在ランク4への所属確率も高いので，潜在ランク3から4に移行しつつあると考えられる（木村，2011）．より具体的に述べるならば，図18の受験者には，「現在のところ5段階中のランク3だが，まだランク2にも近い状況である．易しい問題の中にもできないところがあると思われるので，より難しい問題に取り組む前に，易しい問題の中で不得意なところを復習するとよいだろう」というフィードバックが考えられる．一方，所属ランクは同じでも，図19の受験者には，「現在のところ5段階中のランク3だが，一つ上のランク4にも近い状況である．基礎的な問題はほぼマスターしていると思われるので，より難しい問題に積極的にチャレンジするとよいだろう」というフィードバック

が考えられる．

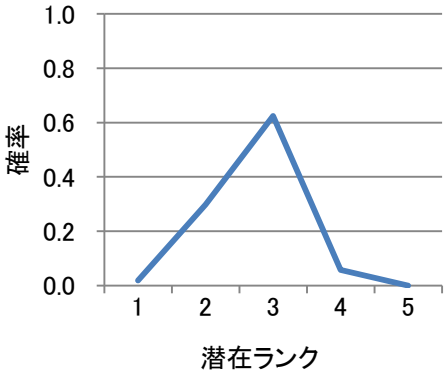

図 18　RMP の例 1

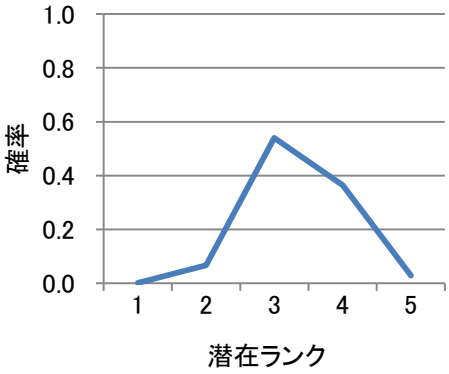

図 19　RMP の例 2

また，同一の学習者のRMPを時系列で追いかけることで，図20（3か月間をあけて受験した2つのテスト2Aと2Gの同一学習者のRMPの変化）のように学力の変化を表現することも可能である（木村，2011）．この場合，RMPの変化する図を受験者に見せて，「3か月前と所属するランクは5段階中の3で変わりないが，今回は前回に比べて易しい問題に着実に正解できるようになっている．今後は，少し難しい問題にもチャレンジして力を伸ばしていくとよいだろう」といったフィードバックを与えることが可能である．

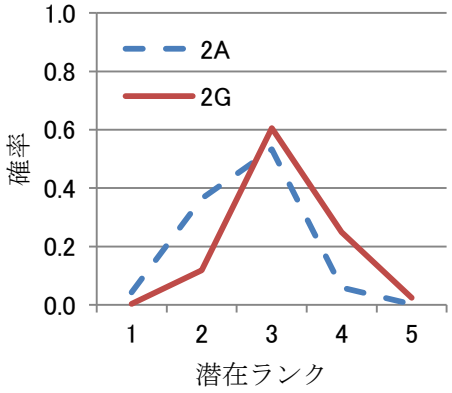

図 20　同一学習者の RMP の変化

1.5.3.　LRT の有用性

LRTは次の4つの側面から有用なテスト標準化理論であると考えられる．実際に教育場面の学習評価でLRTを利用した実践例も，すでにいくつかある．英語語彙テストの分析（小泉・飯村 2010），中学の数学学力テストの分析（松宮・荘島 2008，松宮・荘島 2009），大学生のジェネリ

ックスキルを測定する試み（成田・荘島・宇佐美 2010）などである.

(1) 測定方法論的側面

テストという測定道具は, どのくらい測定対象を細かく区別できるかという意味での解像度が高いとは言えない. 荘島（2008a）は,「テストはそもそも連続的に学力を評価できるほど信頼性が高い測定道具ではなく，10段階くらいにランク付けることがせいぜいである」と述べている. 測定道具としてのテストの解像度の低さを認識し, いくつかの段階に区分することを目指すLRTは測定方法論的側面から理にかなったものである.

(2) 教育心理学的側面

テスト結果を連続的に細かく表現することは, その解像度の低さから測定結果の変動が大きいために，教育心理学的問題を生む. 荘島（2008a）は「連続尺度は，受験生や学生に1点でも多く得点をとろうという受験者心理を助長し,『テストテクニック』のような本来学生たちに求める学力とは異なるような技術が塾や時には学校で教えられている」こと「学力は一昼夜で劇的に変化しないにもかかわらず, 連続尺度の不安定な乱高下で受験者の不安をあおっている」ことを指摘している. したがって，テストデータの分析にLRTを用いることで，いくつかの段階で生徒を評価するようになれば, 少し腰をすえて努力をしないと学力が上の段階に評価されないので，小手先の技術を抑制することに貢献することが期待される.

他方で，宇佐美（2009）の「離散的な潜在ランクの利用により，学習者が一つ上の潜在ランクにステップアップする困難度が平均的に高まることによって, 達成感を持続的に得ることができず，その結果動機づけが低下してしまう学習者がでてくる可能性もある」という指摘もある. しかし，この点については，テストの結果を潜在ランクだけでなく，前述のRMPを使って多義的に示すことである程度回避できるのではないだろうか.

(3) 教育現場の評価体制の側面

教育現場における評価は, 5段階評価に代表されるように, 段階評価で要約されることが多い.「指導要録，通知票，調査書，作品・レポート・実技テスト，学力の文章表現など，教育現場で行われている評価体制は，順序尺度に帰着する」（松宮・荘島，2008）. この点から考えて，LRTは教育現場に即したテスト理論である.

(4) 品質管理・アカウンタビリティの側面

テストの品質管理とアカウンタビリティを向上させる試みは, 様々な場面で行われている. テストの点数や合否を示すだけでなく, そのテストの結果から, 受験者にどのような能力が備わっているかについて説明できることが望ましい. いろいろなテスト理論を使って, テストの結果とCDSの対応づけを行い, 受験者に提示しようとする試みが増えている（日本英語検定協会, 2007; 小山･木村, 2011; 野上･林, 2011）.

松宮・荘島（2009）の「段階評価を導入することにより，段階評価により区別される各能力段階（潜在ランク）の特徴を，Can-Do Chartとの関連で示すことが，連続尺度のもとで検討するよ

りも容易に行える．……テストから作成されたCan-Do Chartはテストの説明資料・学力達成への道標になる」という主張からも，LRTは段階評価のためのテスト理論というだけでなく，CDSの対応づけを行うためのテスト理論といってもよい．換言すれば，LRTは「Can-Do-Chartの作戦支援ツールといっても過言ではない」（荘島, 2010b:108）．

1.6. 本研究で使用するテスト理論

実際に CAT 開発を始める前に，これまでに論じたテスト理論のうち，どの理論に基づいて分析を行い，CAT を開発するかを決める必要がある．本研究では RM と LRT の 2 通りで分析を進めていくことにした．その理由は以下のとおりである．

(1) サンプルサイズの問題

本研究はオープンソースを使った小規模な CAT 開発が目標であり，事前テストにおいて受験協力者として確保できる人数は 200 人前後であった．サンプルサイズがいくつ以上なら，十分な精度で分析できるかは，どのような項目をどのような集団に対して実施するかによって異なるので，一概には論じられないが，大まかな目安として，大友(1996: 98）は，1PLM では 100-200，2PLM では 200-400，3PLM では 1000-2000 というサンプルサイズ目安を示している．したがって，サンプルサイズが 200 程度になる見込みの本研究では，2PLM や 3PLM は候補から除外された．LRT は，分析するランク数によって必要とされるサンプルサイズが異なると考えられる．具体的な目安は示されていないが，これまで試行的に行った分析から，分析するランク数を 5 程度に抑えれば，200 前後のサンプルサイズでも十分分析に耐えられることを確認している．サンプルサイズの制約から，本研究では 1PLM と数理的に同義である RM と LRT によって分析を行うことにした．

(2) テストデータへのアプローチの理念

IRT の 1PLM と RM は数理的にはほぼ同じであるが，1.4 節で説明したようにテストデータへのアプローチの理念に根本的な違いがある．本研究が目指すところは，モデルを作り出すことではなく，客観的な測定結果が得られるようにデータを整理して，意味のある構成概念を作り出すことであるので 1PLM ではなく RM を採用した．

(3) 段階評価と診断的情報の提供

IRT や RM の分析から段階評価をすることも不可能ではないが，結果が連続的に表現される IRT や RM よりも，結果を最初から段階的に表現する LRT の方が，段階評価に適しており，テスト結果を CDS と結びつけたり，なんらかの診断的情報を提供しやすいと考えられる．先述のとおり，LRT の分析での，最低サンプルサイズは，分析するランク数によっても異なり，IRT のように定量的な目安は示されていない．また，LRT における項目除去の方針や CAT アルゴリズムについても先行研究がなく，RM の分析と比較しながら開発を進めていくことが有効であると考え，LRT による分析と RM の分析と並行して行うことにした．

2. コンピュータ適応型テスト（CAT）

テスト理論とコンピュータ技術の発達によって今日の CAT があることは確かであるが，CAT の開発と実施を考える前に，そもそも CAT の根源は何かについて考察し，そのアルゴリズムの基本を整理しておきたい．また，CAT の利点と問題点を整理し，CAT 開発のフレームワークを紹介し，実践的研究を行う上での指針と枠組みを提示する．

2.1. CAT の根源

CAT の根源は何なのであろうか．紙と鉛筆によるテスト（paper-pencil test, PPT）が，コンピュータによるテスト（computer-based test, CBT）になり，それが発展して CAT になったという単純な流れではない．CAT には，コンピュータを使ってテストを実施するという面と，受験者の反応に合わせてアダプティブに出題するという 2 つの面が混在している．言い換えるならば，CAT は，テストのコンピュータ化という側面と，テストの個別化という 2 つの側面を持っている．

コンピュータ化という側面から考えるならば，CBT から発展して CAT が誕生したというとらえ方も間違いではないが，テストの個別化という側面から考えると，集団テストを前提とした PPT や CBT の延長線上にあるのではなく，個別面談テストの延長線上にあると考えられる．つまり，CAT での出題状況は，個別面談テストで，試験者が受験者に質問をし，その反応を見ながら，質問を調整しながら受験者の能力を探る状態に似ている．体系化された個別面接という意味で，ビネーの IQ テスト（Binet’s IQ test: Binet & Simon, 1905）に CAT の根源を見つけることができる．

ビネーの IQ テストの仕組みについて，図 21 を使って大まかに述べる．あらかじめ精神年齢（mental age）ごとに 10 項目からなるテストレットが用意されている．各テストレットはその精神年齢と同じ実年齢の児童が約 50%の確率で正答できる項目からなる．対象者の実年齢を参考に，どの精神年齢のテストレットから実施するかを決めて実施する．図 21 の場合，精神年齢 9 歳のテストレットを実施し，10 項目中 6 項目正解している（項目を実施した順番を表す番号の後ろに，正解の場合は＋不正解の場合はーで表記されている）．最初のテストレット終了後は，シーリングレベル（ceiling level: 正答率 0%になるレベル）とベーサルレベル（basal level: 正答率が 100%になるレベル）が見つかるまで，テストレットをひとつずつ上げていくか，ひとつずつ下げていく．図 21 の場合は先にベーサルレベルを，その後シーリングレベルを見極めている．ビネーの IQ テストの特徴は，いろいろな観点で現代の CAT アルゴリズムに通じるところがある．

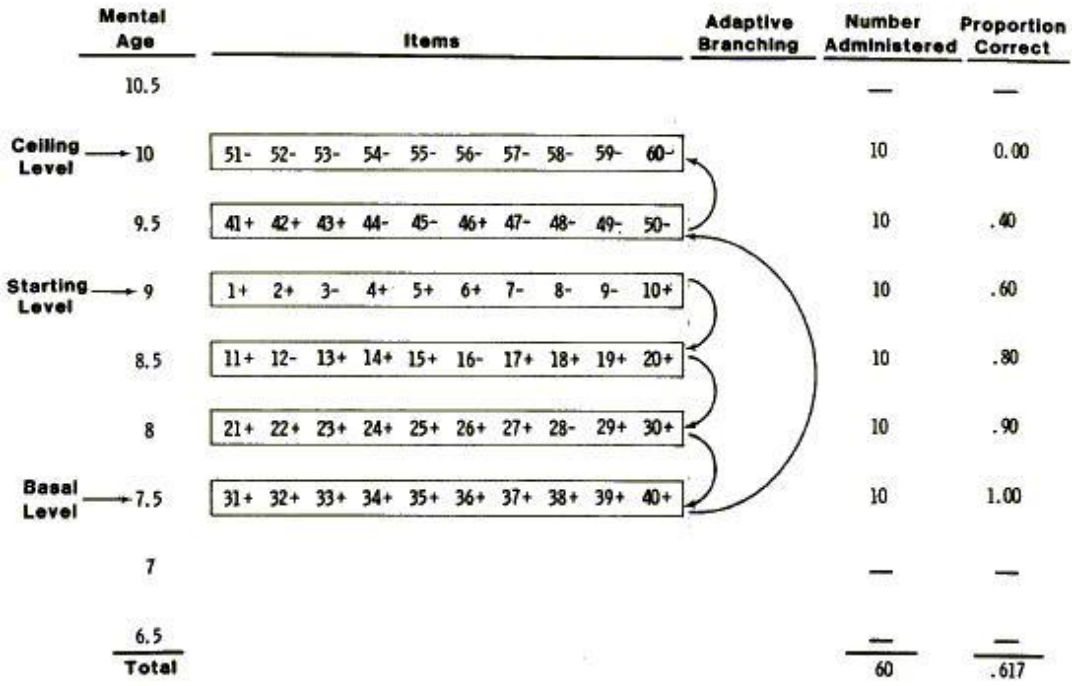

図 21　ビネーのIQテスト実施の流れ [3]

このビネーの IQ テストを，テストレット単位でではなく，項目単位でアダプティブにしたものが，stratified adaptive test (Weiss,1973)である．図 22 は，その実施の流れを示している．1 項目ずつ，正解すればひとつ上の精神年齢の項目を，不正解すればひとつ下の項目を出題してゆく．出題しようと思うレベルの項目がすべて使われている場合は，さらに 1 つ上（または下）のレベルの項目を出題する．そして，1 つの精神年齢用に用意された 10 項目すべてが不正解となった段階で，そこをシーリングレベルにする．図 22 の場合は精神年齢 9 歳を初期レベルとし，1 項目ずつその正誤によって次に出題するレベルを調整し，44 問目で精神年齢 10 歳の項目すべてが不正解となり，シーリングレベルが求められている．IQ 以外の潜在能力を測定する場合でも，このようにレベル（層）ごとに項目を準備しておいて，正答率が 0%になるシーリングレベルと正答率が 100%になるベーサルレベルを求めるという手法は，個に応じた個別テストの手法として有効であり，項目（あるいはテストレット）選択のあり方は，現代の CAT アルゴリズムに通じるものがある．

[3] A Schematic Binet Test Administration: IACAT のホームページ（http://iacat.org/node/442）より引用

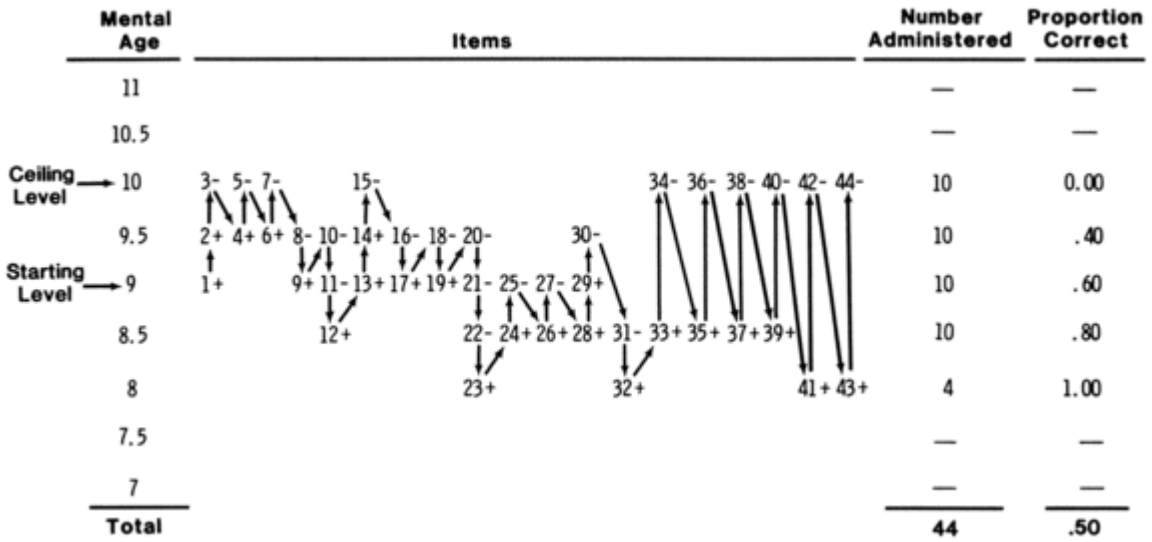

図 22　Stratified adaptive test の流れ[4]

よりはっきりした現代の CAT アルゴリズムの原型は，Lord (1971)の self-scoring flexilevel test に見ることができる．一見，集団に対して一斉に実施される PPT のようであるが，個に応じたテストをする仕組みがテスト用紙自体に組み込まれている. 受験者のレベルにあった問題を解かせ，難しすぎる（あるいは易しすぎる）問題は解かせないことで，精度を落とさずに効率よく受験者の能力を推定するという，現代の CAT の目指していることが PPT で実現されている．実物を目にしないとなかなか分かりにくいが，21 項目を印刷して 11 項目を受験させる場合のテスト用紙のレイアウトが，Lord (1980)に図 23 のように図示されている．

出題する奇数個の問題の困難度があらかじめ大まかに判断されており, ちょうど真ん中の困難度の問題がテスト用紙の一番上中央に初期項目として印刷されている. 初期項目より易しい問題がテスト用紙の左側に難しい順に，初期項目より難しい問題が易しい順に印刷されている．左側の問題は赤字で問題番号が 1 から振られており, 右側の問題は青字で問題番号が 1 から振られている．すべての問題は多肢選択式で，解答をマークするたびに，赤か青の色が現れる仕組みになっている．受験者はその色に従って，赤が出た場合は，左側の次の番号の問題を，青が出た場合は，右側の番号の問題を解く．受験者は最初の問題を除く設問の半分だけを解くことになる．

この Lord(1971)の self-scoring flexilevel test の手法をそのまま CAT 化することもできる．アイテムバンクに項目数が少なく，受験させたい項目数の 2～5 倍程度しかない場合は，単純だが有効な方法かもしれない．実際，De Ayala & Koch(1986)は，Lord(1971)の方法をコンピュータで実現し，シミュレーションデータにより，flexilevel CAT が，ベイズ推定法に基づく CAT の結果と比較して遜色ないことを示し，IRT に基づき項目特性を求められたアイテムバンクを用意しなくても実行可能であり，教室環境で有効な方法であることを示唆している．

[4] Weiss (1985)より引用

真ん中の困難度の項目(困難度 11 位) ① ② ③ ④	
1.* 少し易しい項目（困難度 12 位） ① ② ③ ④	1.† 少し難しい項目（困難度 10 位） ① ② ③ ④
2.* 少し易しい項目（困難度 13 位） ① ② ③ ④	2.† 少し難しい項目（困難度 9 位） ① ② ③ ④
3.* ・ ・ ・ [より易しい項目] ・ ・	3.† ・ ・ ・ [より難しい項目] ・ ・
10.* 最も易しい項目（困難度 21 位） ① ② ③ ④	10.† 最も難しい項目（困難度 1 位） ① ② ③ ④

*数字が赤で印字されている　　†数字が青で印字されている

① ② ③ ④は選択肢でマークする（削ると）赤または青色が現れる

図 23　21 項目を印刷したflexilevel testのレイアウト [5]

2.2. CAT のアルゴリズム

CAT に限らず，あらゆるテストは，そのアルゴリズムは次の３つの部分からなる ― 1. How to START, 2. How to CONTINUE, 3. How to STOP（Thissen & Mislevy, 2000）. 3.2 の RM に基づき CAT を実装するプログラム UCAT（Linacre, 1987）のアルゴリズムの説明と，4.2 の LRT のアルゴリズムの提案では，この３つの部分に分けて説明を行う. CAT 全体の流れを理解するために，CAT のアルゴリズムをもう少し細かくフローチャートの形で書くと，図 24 のようになる.

[5] Lord (1980)をもとに Thissen & Mislevy (2000: 103)が作成した図を翻訳加筆した.

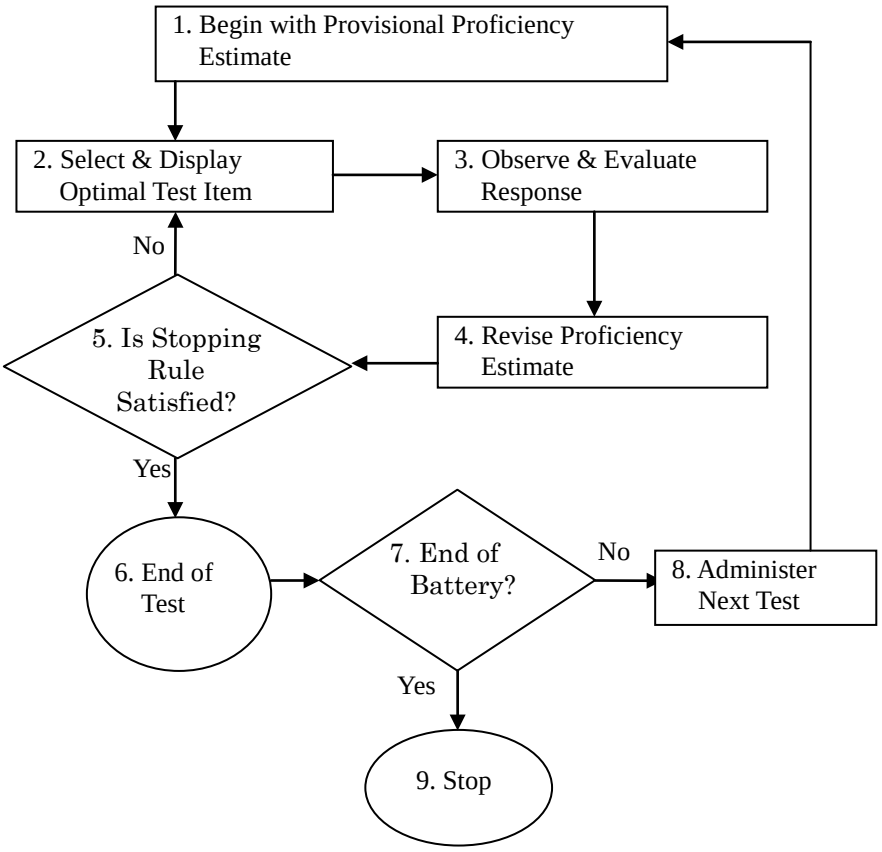

図 24　CATアルゴリズムのフローチャート [6]

これまでに，CAT のアルゴリズムは数多く提案されており，細かく条件を変えてシミュレーションが行われたり，実際に CAT を実施して研究が行われている．その多くは IRT に基づくものである．LRT に基づくアルゴリズムを提案する前に、ここでは，Halkitis（1993）が看護学の学生の薬理学の能力を測定する CAT として開発した RM による CAT アルゴリズム(図 25)を参考に，現代の CAT が一般的にどのような流れで行われているのか，工夫すべき点はどこにあるのかについて，概略をつかむことにする．特に注目に値する工夫がなされている箇所は，図中に丸数字を付した．

[6] Thissen & Mislevy (2000:106）より引用

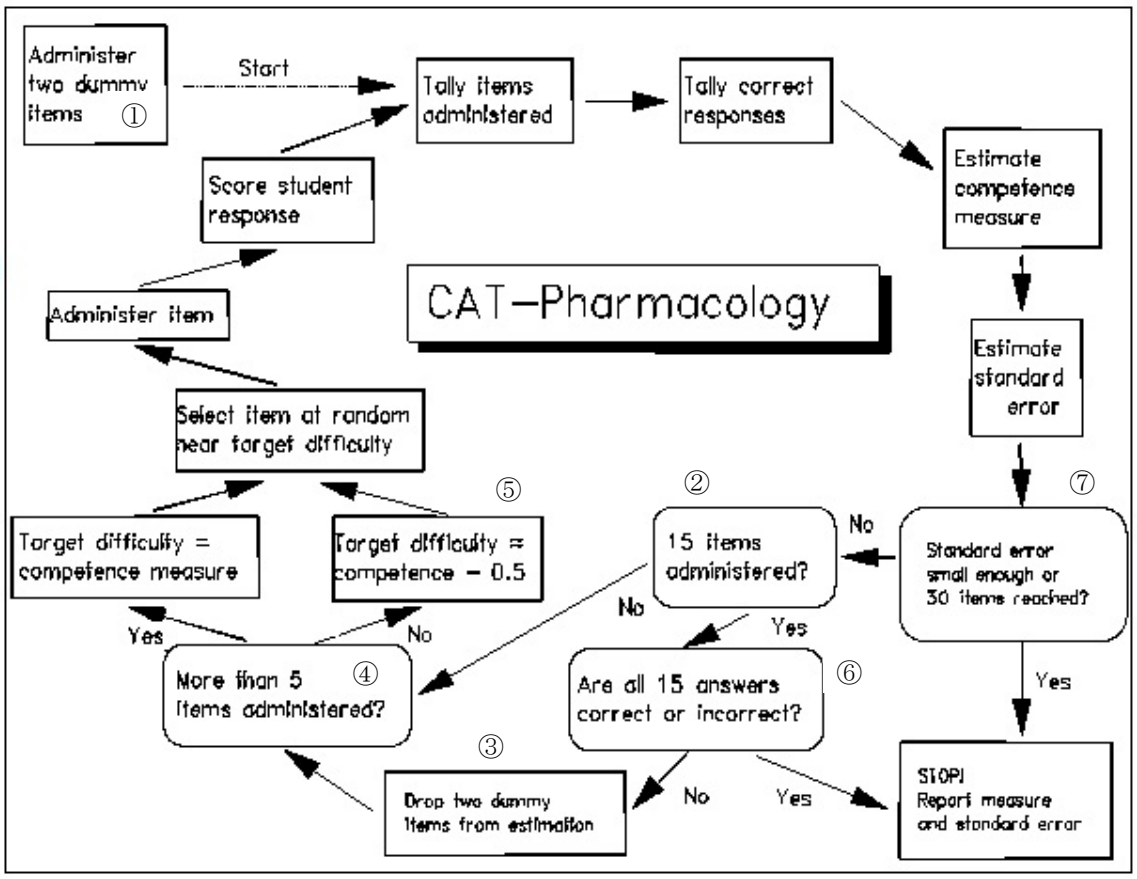

図 25　CAT-Pharmacology algorithm[7]

①～③「CAT初期段階の偶然正解やケアレスミスによる推定バイアスの軽減」: 初めにダミー項目を2つ置いている. 項目困難度が0のものを1問正解・1問不正解したものとして初期能力値を推定し，実際に15項目解答が終わるまで，毎回の能力推定値の計算にもこのダミーを含め，15項目解答が終わったらダミーを除外している. これにより初期の緊張による失敗や偶然の正解により能力推定が大きく変わってしまうことを防いでいる．これは，attenuation paradox （Lord & Novick, 1968）として知られている事象（実際の能力が初期能力推定値からずれていた場合，その後の能力推定と項目選択が不適切に行われてしまうこと）への対処である．m項目受験した後の能力推定は，直前の能力推定値をθ_m，正答数をR_m，能力θ_mの受験者がj番目の項目に正答する確率を$P_j(\theta_m)$とし，

$$\boldsymbol{\theta_{m+1} = \theta_m + \frac{R_m - \sum_{j=1}^{m} P_j(\theta_m)}{\sum_{j=1}^{m} P_j(\theta_m)\left(1 - P_j(\theta_m)\right)}} \tag{23}$$

[7] Halkitis (1993)より引用（図中の丸数字は筆者が説明のために加筆した）

で表される．なお，$P_j(\theta_m)$は式(13)で定義されている$P_j(\theta_i)$と同義.

④・⑤「受験者の心理的負担軽減（項目困難度の調整）とitem exposureの調整」：第5項目までは，推定能力よりも0.5logits低い困難度の項目の近くからランダムに選んでいる．推定能力と同じ困難度の項目を選ぶと正答確率が50%になるのに対して，この場合62%になり，CATに不慣れな学生に対して心理的負担を少なくしている．また，ランダムに選択することで項目の使用頻度（item exposure）をコントロールしている．

⑥・⑦「終了条件の定義に関する工夫」：15項目すべて正解あるいは不正解の場合テストを終了し最高あるいは最低スコアを提示している．それ以外の場合は30項目解答するか，SEが0.4logitsより小さくなった場合に終了させている．さほど大きくないアイテムバンクの場合，平均的困難度の項目からスタートして15項目すべて正解あるいは不正解の場合，それ以上その受験者の推定能力に適した項目があるとは考えにくく，最高あるいは最低スコアとするほかない．測定の精度を重視するならばSEのみを使った終了条件も考えられるが，受験者が間違った解答行動を取ると，指定したSEになかなか達しないことがある．終了条件に上限項目数を設けておくことは，極めて現実的であり，よく使われる策である．ちなみに，m項目受験した後の推定のSEは，

$$\boldsymbol{SE_{m+1}} = \sqrt{\frac{\mathbf{1}}{\sum_{j=1}^{m} \boldsymbol{P_j(\theta_m)}\left(\mathbf{1} - \boldsymbol{P_j(\theta_m)}\right)}} \tag{24}$$

で計算される．

2.3. CAT の利点

CAT の利点は大規模・集団テストのことを想定して議論されることが多い．全受験者が同じ項目を受験する PPT や CBT と比較して，CAT の利点は多くの場合，次の５つに大別して語られる（Weiss, 1982; Green, 1983; Zickar et al.: 1999）.

（１） セキュリティーの向上（increased test security）
（２） 得点の即時報告（instantaneous score reporting）
（３） 多用な項目形式（variety of item type）
（４） 効率の強化（enhanced efficiency）
（５） 優れた測定精度（superior measurement）

最初の３つは物理的な利点であり，CBT にも当てはまるものであり，PPT と比較した場合の CBT と CAT の利点である．（１）と（２）は PPT で行われていたテストを CBT または CAT に移行させるだけで実現される利点であり，それによる経済効果の大きさも指摘されることがある．一方，（３）については，テストを実施する媒体が紙面でなくコンピュータになったことで生まれた利点である．項目刺激提示のマルチメディア化（文字や静止画像だけでなく，音声や動画の活用など）や応答収集の多様化（単純な多肢選択以外にも数値・文字入力や音声録音による応答

など）のことである．これらについては，早くから指摘されているが，実用化された例はあまり多くない．最後の２つは測定論的利点であり，これは CAT だけに当てはまる利点である．全員がすべて同じ項目を解答する PPT や CBT に比べて，問題数を少なくしても測定の精度が下がらないということである．「CAT にすることによって，項目数を半分にしても測定の精度は変わらない」ということが，しばしば強調されるのはこの２つの利点によるものである．確かに，心理測定論的見地からだけテストのことを考えるなら，これは CAT の大きな利点であるが，受験者の心理学的側面を考えると，全面的に CAT の利点と言いがたい面もあるので，次節でさらに考察を加える．

2.4. CAT の問題点

CAT のかかえている問題点については，CAT をどの観点から捉えるかによって異なるが，次の４種類に分類される．

（１） 事前テスト実施と分析・アイテムバンク管理の手間
（２） 受験におけるコンピュータ操作と解答方法に関する不安
（３） 全受験者が同じ項目を解かないことに対する不公平感
（４） 実施項目中の正答数割合の低さによる心理学的影響

まず，（１）はテストを準備・実施する立場からの問題点である．CAT でなくとも，しっかりと標準化されたテストを用意するには，事前テストの実施と分析は不可欠である．しかし，CAT の効率性と測定精度を高くするためには，用意すべき項目数は，アダプティブでないテストの場合よりも多くなる．また，継続的に同じアイテムバンクで CAT を実施した場合，項目間でその使用頻度にばらつきが出てしまうので，item exposure を管理しつつ，適切な時期にアイテムバンクに新しい項目を追加していく作業が欠かせない．

２番目の問題は，さらにコンピュータ・リテラシーの問題（コンピュータ操作が不慣れであること）と，CAT でないテストの場合と異なる解答行動を要求される問題（出題される順に解答し，後から解答を変更することができないこと）に分けられる．前者は，コンピュータの普及とともに全体としては問題にならない程度になってきているが，ごく少数ではあっても，コンピュータに対する不安からテスト結果に影響を与えてしまうケースがあることは無視できない．後者は，CAT のアルゴリズム上，回避することは難しい問題である．CAT というテスト方法の普及が進むことで，受験者がこの解答行動に慣れていくことで，問題視されなくなることが期待される．

３番目の問題は，２番目の後者の問題と同様，CAT のアルゴリズ上，不可避なことである．受験者に対して，CAT の仕組み（出題方法と採点方法）をよく説明した上で CAT を実施することが大切である．この２つの問題は，受験者に CAT に対する理解を深めてもらうことで，解決を図るしかないであろう．

最後の問題は，CAT を受験することで，学習への動機づけや自己効力感を低下させる可能性が指摘されているので，改善すべき問題だと考える．以下，これまでの研究を振り返りながら，

改めて問題を整理し，解決の方向性を考えてみたい.

一般的に多くのCATのアルゴリズムにおいて, 測定している受験者に対して最も情報量の多い（精度の高い）項目を中心に選ぼうとする．つまり，それは受験者にとっては，自分の能力レベルに関わらず, 正答確率50%前後の項目ばかりが出題されることを意味する. 学校教育においても，資格試験においても，実施されるテストは，ほとんどの場合，平均点が100点満点で，60～70点ぐらい（場合によっては80点）の正答率になるように作成される．資格試験で50%の正答率で合格するように設計されるということはまずなく, 大学入試センターの問題もほとんどの科目が平均点は60ぐらいである．日本の中学や高校で実施される定期試験などは，多くの場合平均は70～80点ぐらいである．CATによるテストを経験したことがない受験者が，CATの仕組みを十分理解できないまま受験すると，心理的に落ち込んでしまう．CATにおいて正答確率を50%に設置することは, 受験者のテストに対する動機づけを維持させるには低すぎるかもしれないとう指摘もある（Andrich, 1995）. Gershon (1992)は,「CATの最初の項目ならびにおそらくすべての項目は，受験者に達成感を持たせるために，少し易しい問題にすべきではないか」と指摘している.

Bergstom et al (1992)は，726 項目からなるアイテムバンクを使い，予め設けた基準値（.15）を超えているか否かを判定する CAT（終了条件：能力推定値が基準値より SEM の 1.3 倍上か下となるか実施項目が 240 を超えた場合）において，正答確率が 50%，60%，70%の項目が選ばれるように項目選択ルールを変化させて 225 人に対して実施し，結果を比較した．その結果，正答確率が 50%ではなく，60%や 70%のものが選ばれる CAT であっても，50%の場合と同じ SE で終了するまで，実施項目がわずかに増えるだけであることを示した．Tonidandel & Quiñones (2000) も同様の結論に達している.

Ponsoda et al (1999) や Tonidandel et al (2002) は，同様に異なる正答確率で項目選択が行われる CAT を実施するとともに，受験者の CAT に対する反応についても調査し，受験者の動機付けの観点から見ると，情報量を重視して正答確率が 50%の項目が選択される CAT よりも，より易しめの（正答確率が高い）項目が選択される CAT の方が望ましいことを示している.

易しい項目を選択すると項目数が増え，テスト時間短縮という CAT の利点が損なわれるという懸念がある．しかし，一般的にテストにおいて,「易しい問題は難しい問題よりも時間がかからない」（false＞correct phenomenon）ということを考慮すると，少し項目数が増えても項目の困難度が下がるなら，テスト時間という意味では大差ないという考え方もできる．Hornke (1995, 2000) は，テストの解答時間を調査し，解答時間は正答確率と直接的な関連があり，実際にこの false＞correct phenomenon が存在することを実証している.

CAT における項目の困難度の調整は，当然のことながら，全項目でなく部分的に行うことも可能である. Häusler & Sommer (2008) は, 126 項目のアイテムバンクから 20 項目を出題する CAT で正答確率について次の 6 とおりを用意し，シミュレーション研究を行った．１）全項目 50%，２）全項目 60%，３）全項目 70%，４）全項目 80%，５）ランダムな位置で 4 分の 1 だけ 80%他は 50%，６）ランダムな位置で 2 分の 1 だけ 80%他は 50%，その結果，１）から２）に正答率を変えても信頼性はそれほど落ちないが（.801 から.771），３）や４）にすると，信頼性が大

きく下がり（.723 と.624），平均より上の潜在能力の場合に特にバイアスが大きくなってしまうことを示した. 一方，５）や６）のようにテスト全体ではなく部分的に困難度を調整した場合は，明らかに易しい問題を出題しても，測定精度はそれほど落ちない（信頼係数で.779 と.775）ことを示した．Bergstrom et al (1992)や Tonidandel & Quiñones (2000)に異なり，正答確率を 70%のものが選ばれるようにすると CAT の信頼性が大きく低くなっているのは，アイテムバンクの項目数が小さいことが大きく影響していると思われる.

さらに Häusler & Sommer (2008) は，１）と５）と６）の条件で実テストを行い，同時に解答に対する自信についても調査した. 3 条件間に受験者の能力推定において統計的に有意な差は見られなかったが，受験者の自信については１）と５）並びに１）６）の間で，統計的に有意な差が見られた（ただし，５）と６）の間に統計的に有意な差は見られなかった）. 以上２つの実験から，受験者の心理的側面を考えて，項目困難度を下げる場合は，すべての項目を正答率 70%に下げるより，4 分の 1 や 2 分の 1 だけ明らかに易しい項目（正答確率 80%）にする方が，測定の精度を落とさず，かつ，受験者の自信も下がらないことが示唆された．また，受験者の解答に対する自信は，明らかに易しい項目（正答確率 80%）の割合が 4 分の 1 でも 2 分の 1 でも，大きく変わらないので，全体の 4 分の 1 に，明らかに易しい項目（正答確率 80%）を出題するのが，測定精度と受験者の心理的側面の両方に配慮した結果が得られることが示唆された.

Häusler & Sommer (2008)では，易しい問題を出題する位置はランダムであるが，CAT でないテスト（同じ項目を全受験者に解答させるテスト）の場合，多くの場合，最初に易し目の問題を，後の方に難し目の問題を配置することが多い. 最初に易しい項目を出題することに関する効果については，その後のテストのできに影響しないという研究（Lunz & Bergstrom, 1994）と，受験者の情意面と動機づけにプラスの効果があるという研究（Mills, 1999）の両方がある．CAT における項目困難度の調整は，測定の精度を優先するなら，正答確率 50%近辺を選ぶのが最適であるのは明らかであるが，受験者の心理学的側面（動機づけや自信や自己効力感など）へ配慮する場合，困難度の設定，複数の困難度を設定する場合の割合と位置など，まだまだ明らかにしなければいけないことが多々ある．明らかなことは，項目困難度の調整を媒介として，CAT の精度と受験者の心理的側面の間にはトレードオフの関係が存在するということである.

2.5. CAT 開発フレームワーク

CAT を開発するには，テスト理論やアルゴリズムのことだけでなく，開発全体の流れとそれぞれの段階で行うべきことや注意すべきことを把握しておくべきである．そのためには，CAT 開発のフレームワークを持った上で，開発を始めるのがよい．CAT 開発の段階を追って整理したフレームワークとして，Thompson & Weiss (2011)のもの（表 2）がある．極めて実践的な立場から CAT 開発の段階が整理されており，大変参考になる．本論文の後半，実践編では，このフレームワークに沿って，これまで行ってきた実践的研究を整理する.

第1段階は，導入しようとする評価の場面においてCATを導入することが望ましいのか，それに注がれる労力と費用を考えて価値あるものかを, CATのためのアイテムバンクを構築し始める

前に，判断しようとするものである．どのようなモデルとアルゴリズムを使ったCATを開発するかの検討も必要である．モデルで使用するパラメータの数が多くなればなるほど，第3段階で行われる事前テストで必要とされる受験者数は多くなるので，どの程度の規模の事前テストを実施可能かによって，モデルを選択することも重要である．その上で，どのような特性の項目をどのくらい用意すれば，どの程度の精度のCATを開発できるのか，モンテカルロ・シミュレーションで確認することが推奨されている．これから作ろうとするアイテムバンクの項目特性の分布を想定し，測定しようとする受験者集団の能力分布を想定し，そのサイズを仮定して架空のアイテムバンクと受験集団を生成し，評価の場面で必要とされる精度に達するには，何項目出題する必要があるのか，想定したアイテムバンクのサイズは十分かなどを，何通りかのシミュレーションを実施して見極めることになる．つまり，架空のデータでこれから開発するCATのアイテムバンクの青写真（blueprint）を描いてみるわけである（Veldkamp & van der Linden, 2010）．

表 2　CAT 開発のフレームワーク

（Thompson & Weiss (2011)の Table 1: Proposed CAT framework を翻訳）

ステップ	段階	中心課題
1	実現可能性と適用性の評価 計画調査の段階	モンテカルロ・シミュレーション，投資対効果検討評価
2	アイテムバンクの項目作成，あるいは既存のバンクの活用段階	項目作成，見直し
3	事前テストの実施とアイテムバンクの項目特性を分析する段階	事前テスト，項目分析
4	最終的な CAT の仕様を決定する段階	事後シミュレーションまたはハイブリッド・シミュレーション
5	実際に CAT を世に出す段階	出版と配布：ソフト開発

第2段階は，項目を作成しアイテムバンクを構築する段階である．既存のアイテムバンクを援用して付け加えることも考えられる．第1段階のアイテムバンクの青写真にそって項目を作成していくべきである．項目の困難度や識別力だけでなく，コンテンツのバランスも考慮しながら，項目作成は行われるべきである．そのためには，テスト理論専門家とテスト内容の専門家が協力して作業に当たる必要がある．

第3の段階は，第2段階で用意されたアイテムの項目特性を測定するために事前テストを行う段階である．第1段階で検討したテスト理論（モデル）によって，事前テストで必要とされる受験者の数は異なる．また，全項目を同じ受験者に一度に実施することは，よほど小さなアイテムバンクを目指す場合でない限り，不可能である．そのため，何回かに分けて事前テストを実施することになるので，それらの結果を後で一つにまとめられるように（等化できるようにするために），

計画的に複数の事前テストを用意する必要がある．事前テストで集められたデータを分析して，まず確認すべきことは，それらの項目が同じ能力を図っているのか（一元性があるか）を確かめる必要がある（multidimensional modelを利用する場合を除く）．選択したモデルにフィットしない項目は，アイテムバンクから除去する必要がある．これらの作業を終えた段階で，青写真とどの程度近いアイテムバンクが構築されているかを確認し，不足している部分が多い場合は，追加の項目作成と事前テストを実施する必要がある．青写真通りのアイテムバンクがすぐに完成することは稀であり，追加の項目作成や事前テストが困難な場合もある．その場合は，次の第4段階でCATの仕様を変更するのも現実的な選択であろう．

第4段階では，完成したアイテムバンクでのシミュレーションを行い，最終的なCATの仕様を決定する段階である．ここでいうCATの仕様とは，１）初期推定能力の決め方（starting point），２）項目選択アルゴリズム（item selection algorithm），３）採点方法（scoring method），４）終了条件（termination criterion）のことである．第1段階のシミュレーションと異なり，この段階のシミュレーションは，実際のアイテムバンクを使って行われる．アイテムバンクの規模が小さい場合は，実際の項目特性と実際に事前テストで解答した受験者の応答を組み合わせてシミュレーションすることも可能だが，大きなアイテムバンクが構築された場合には，ある受験者が解答した項目の割合は少ないので，実際の応答を使えないケースが多くなる．そのような場合は，事前テストで解答していない項目については第1段階と同じ方法で応答を生成し，シミュレーションを行う（Nydick & Weiss, 2009）．すべて事前テストの解答結果をもとにシミュレーションを行う前者の方式は，事後シュミレーション（post-hoc simulation）あるいは実データシミュレーション（real-data simulation）と呼ばれ，事前テストの解答が使える部分はそれを使い，事前テストでの解答がない部分は解答をモンテカルロ・シミュレーションによって生成して行う後者の方法は，ハイブリッド・シミュレーション（hybrid simulation）と呼ばれる．

3. ラッシュモデルに基づく CAT（RM-CAT）

RM に基づく CAT（RM-CAT）のアルゴリズムは 1980 年代から提案されており，実践的研究も多く行われている．2.2 で紹介した Halkitis（1993）の薬理学の CAT もその一例である．本章では，Wright (1988)を元に RM-CAT のアルゴリズムの中核を整理するとともに，筆者らが Moodle 上で CAT を実装するためのシステムとして開発した Moodle UCAT の原型である UCAT(Linacre, 2000)のアルゴリズムを紹介する．

3.1. RM-CAT アルゴリズム

RM-CATアルゴリズムの中核となる要素については，Wright (1988)が0から20のステップで整理している．このCATアルゴリズムはある基準となる能力推定値（T）を超えているか否かを判定するためものであり，アルゴリズムも計算方法も非常にシンプルである．受験項目数をL，正解数をR，不正解数をW，項目困難度をDとすると，能力推定値Bは，

$$\boldsymbol{B = \frac{\sum D}{L} + \log\frac{R}{W}} \tag{25}$$

で計算され，*SE*は

$$\boldsymbol{SE = \frac{L}{R * W}} \tag{26}$$

で計算される．項目選択ルールも，正解の後は直前の*D*から2/*L*を加えた困難度の項目を，誤答の後は直前の*D*から2/*L*を引いた困難度の項目を出題するというシンプルなものである．そして，判定は，$T-SE<B<T+SE$ならテスト続行，$B-SE>T$ なら合格，$B+SE<T$ なら不合格とするものである．RMに基づいたより精密な能力推定とSEの計算は(23)式と(24)式で行えるわけだが，簡易方法として，小規模なアイテムバンクで学習課程の習熟チェックを目的とするような場合は有効な手段だといえる．

十分な大きさのアイテムバンクを構築し，継続的にCATを実践していこうとすると，次の3つの問題に直面する．

（1） アイテムバンクを構築するために多くの事前テストを実施しなければいけない

（2） 既存のアイテムバンクに新しい項目を追加していかなければならない

（3） アイテムバンクの項目困難度を再計算しなければならない

長い期間利用するCATのアイテムバンクとして十分なサイズになるまで，事前テストを行えば，その後は新しい項目を追加しなくてもよいなら，とにかく十分に多くの事前テストを実施してからCATをスタートさせるという方法もよいだろう．しかしながら，繰り返しCATを実施していくと，事前テストで推定した項目困難度は，少しずつずれてくることが考えられる．また，ある使用頻度を超えた項目は新しい項目に少しずつ交換しなければいけない状況になる．つまり，CATを継続的に実践していくと，必ず項目の追加，項目困難度の再計算は行わなければいけない．項目追加と推定値の再計算の問題を解決することを目指して作成されたのが，次に紹介するRMに基づきCATを実装するプログラムUCAT (Linacre, 1987)である．

3.2. RM-CAT を実装するプログラムの先行例：UCAT

UCATのUはuseful（有用な）の頭文字であり，Linacre (2000: 15)では，論文中の見出しにおいて，UCAT: CAT with Item Bank Recalibration と表記している．つまり，UCATは「アイテムバンク再計算機能付きの有用なCAT」という意図で開発されたものであることがわかる．アイテムバンクにいくつか項目特性が不確かな項目が含まれていても，RMによる測定が台無しになることはなく（Wright & Douglas, 1975; Yao, 1991），後からアイテムバンク全体の項目区制は再計算することができるという考え方である．再計算の際は，それまでのテスト結果への影響を最小限にと

どめるために，それまでに測定された推定能力の平均値は変更させない（Linacre, 1987）．

UCATでは，logit単位を10倍して100を加え，受験者に結果を報告する時の単位（unit）としている．ただし，以下のアルゴリズムの説明では，理論を理解しやすいようにlogit単位のままで説明を加える．UCATのアルゴリズムを，(1) How to START, (2) How to CONTINUE, (3) How to STOPの3段階に分けて整理すると次のようになる．

(1) How to START

初期能力推定（θ_0）はアイテムバンクに用意された項目の項目困難度の平均値（$AVG(D)$）から，ランダムに0から0.5を引いた値を用いる．すなわち，

$$\boldsymbol{\theta_0 = AVG(D) - 0.5 * RND} \tag{27}$$

ただし，RNDは 0 と 1 の間の値をランダムに発生させた数値

によって決定し，それに基づき，初期項目は $\theta_0 \pm 0.5$ の範囲の困難度の中からランダムに選択される．

(2) How to CONTINUE

初期項目への解答が得られた後は，暫定能力推定値とSEが(23)式と(24)式で計算される．その後は，次に示す下限（lower limit: *LL*）と上限（upper limit: *UL*）の間の範囲から，ランダムに次の項目が選ばれて実施される．もし，この範囲の困難度の項目がアイテムバンクに存在しない場合は，この範囲から最も近い値の困難度を持つ項目が次の項目として選択される．*m*項目終了後の*LL*は，

$$\boldsymbol{LL = \theta_m + \frac{R_{m-1} - \sum_{j=1}^{m} P_j(\theta_m)}{\sum_{j=1}^{m} P_j(\theta_m)\left(1 - P_j(\theta_m)\right)}} \tag{28}$$

によって計算される．これは，*m*項目が誤答であった場合の，*m*項目終了後暫定能力推定値である．*UL*は，

$$\boldsymbol{UL = LL + \frac{1}{\sum_{j=1}^{m} P_j(\theta_m)\left(1 - P_j(\theta_m)\right)}} \tag{29}$$

によって計算される．その後，次に述べる終了条件を満たすまで，解答を得るごとに暫定能力推定値とSEの計算が(23)式と(24)式で繰り返される．

(3) How to STOP

UCATは終了条件として，「指定した項目数に達するまで」，「指定したSE未満になるまで」，「すべての項目を実施するまで」の3とおりの中から選択できるようになっている．

3.3. RM-CAT 実装プログラム UCAT の改良：Moodle UCAT の開発

前節で紹介した RM-CAT を実装するプログラム UCAT は，1980 年代の BASIC で書かれたプログラムであるため，いくつかの方法を試みたが，現代のパソコンでそれを実行することはできなかった．BASIC で書かれたプログラムのソースは，Linacre (2000: 46-58)に公開されているので，それを現在世界中で最も広く使われているオープンソースの LMS である Moodle で稼働するように PHP で書き直し，その一部に改良を加えて開発したものが，ここで紹介する Moodle UCAT である. 当初 Moodle のバージョン 2.0 に合わせて開発を始めたが(Kimura & Ohnishi, 2011)，現在はバージョン 2.3 で開発を続けている(Kimura, Ohnishi, & Nagaoka, 2012)．他にも Moodle 上で RM に基づいた CAT を実装する試みは，Koyama & Akiyama (2011) があるが， Moodle バージョン 1.9 用に開発されたもので，そのソースは公開されていない.

Moodle UCAT を開発するにあたり, UCAT に改良を加えた部分は, 項目選択ルールに Logit Bias という機能を追加できるようにしたことである．Logit Bias に指定する数値を *Bias* とすると，項目選択のために（28）式で定義した *LL* の値は次式のように修正される.

$$\boldsymbol{LL_{biased} = \theta_m + \frac{R_{m-1} - \sum_{j=1}^{m} P_j(\theta_m)}{\sum_{j=1}^{m} P_j(\theta_m)\left(1 - P_j(\theta_m)\right)} + Bias} \tag{30}$$

この機能は選択される項目の正答確率を調整するもので，Logit Bias に正の数値を入れると，選択される項目の困難度は高くなり, 負の数値を入れると, 選択される項目の困難度は低くなる. 正答確率という観点からいうと，正の数を入れると，正答率がより低い項目が，負の数を入れると，正答確率がより高い項目が，選ばれることになる.

他の多くの CAT アルゴリズムと同様，UCAT の項目選択ルールでは，最も多くの情報量が得られるように（*SE* がもっとも小さくなるように），(11)式で説明した IIF が最大となる項目（正答確率が 50%の項目）を選ぶので，受験者はテスト全体を通して 50%しか正解できない状況になる．Logit Bias という機能を追加した理由は，CAT を受験することによって，受験者が学習に対する自己効力感や動機づけを低めてしまう可能性を，抑えたいと考えたからである．Logit Bias と選択される項目の正答確率は，表 3 のような関係にあることがわかっている. これを使って，たとえば, 実施する CAT で 60%の正答確率の問題を出題したいなら, Moodle UCAT の Logit Bias の欄に-0.4 を入力することで実行される.

ところで，RM において，項目困難度（*b*）と能力推定値（*θ*）と正答確率（*P*）の関係は，（13）式に示されている．（13）式を次のように変形させることで，表に頼らず，任意の出題したい正答確率（*P*）から指定すべき Logit Bias すなわち*b*—*θ* の値を求めることもできる.

$$\boldsymbol{\text{Logit Bias} = -\log_e\left(\frac{P}{1-P}\right)} \tag{31}$$

表 3　Logit Bias と正答確率の関係

Logit Bias	正答確率	Logit Bias	正答確率
-4.0	98%	4.0	2%
-3.0	95%	3.0	5%
-2.2	90%	2.2	10%
-2.0	88%	2.0	12%
-1.4	80%	1.4	20%
-1.1	75%	1.1	25%
-1.0	73%	1.0	27%
-0.8	69%	0.8	31%
-0.5	62%	0.5	38%
-0.4	60%	0.4	40%
-0.2	55%	0.2	45%
-0.1	52%	0.1	48%
0.0	50%	0.0	50%

4.　潜在ランク理論に基づく CAT（LRT-CAT）

LRT の特徴と有用性については，1.5.2 で詳しく述べたが，LRT は Shojima（2007a）によって発表された新しいテスト理論であるため，LRT に基づいた CAT 開発と実践についての先行研究はない．モデル適合度指標や情報量規準については，Shojima(2008）によって提案されており，項目の特性を要約する指標についても，熊谷（2007）によって提案されているが，CAT のアイテムバンク構築において望ましくないと考えられる項目を除去する方針は確立されていない. また，LRT に基づいた項目選択ルールや終了条件など，CAT アルゴリズムの提案の先例もない. 本章では，木村・永岡（2012b）と木村・永岡（2011a, 2012a）を元に，LRT における項目除去方針と LRT に基づいた CAT アルゴリズムの提案を行う.

4.1.　LRT-CAT のための項目除去方針の提案

CAT を実現するためには，項目を精査しながら，等化作業を行い，アイテムバンクを拡充していく必要がある．これまで LRT による CAT を実装するためにアイテムバンクを拡充する際に，項目の精査と除去には RM の指標（*Outfit MNSQ* と *Zstd* (*t*-value)）や CTT の指標（IT 相関）を使ってきたが，異なる理論のもとで行う misfit 除去は，当然のことながら LRT の枠組みでの分析と相いれないところがあり，矛盾を生じてしまう可能性がある.

LRT において，項目の特性は IRP で表される．IRP はその項目を受験した場合，各潜在ランクの受験者の正答確率をまとめたもので，グラフ化することで項目の特性を把握しやすい．こ

れは IRT の ICC と似ているところが多い．IRP の形状を見て，望ましくない形を除去するということは可能だが，視覚だけに頼った評価は煩雑になり判断を誤る可能性もある．

数値として IRP 形状を要約しているものとして IRP 指標が提案されている．木村・永岡（2012b）は，IRP 指標 γ と c と a の値をもとに，LRT に基づく CAT（LRT-CAT）のためのアイテムバンク構築において望ましくない項目除去方針を提案するとともに，これまで行っていた RM の指標により除去された項目を比較し考察を加えた．

LRT-CAT のアイテムバンク構築において望ましくないと考えられる項目の IRP は，IRP 指標から見て次のような場合である．

（1）γ=1 の場合

（2）c の値が大きい場合

（3）a の値が小さい場合

たとえば，図 26 のように IRP が常に右肩下がりの場合，γ=1 となる．全体的な下がり具合は c によって判断される．分析した項目の中に複数 γ=1 となる項目がある場合は，c の値がより大きいものから除去すべきであろう．

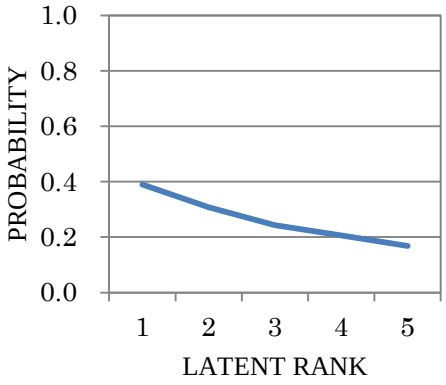

図 26　IRP の例 1

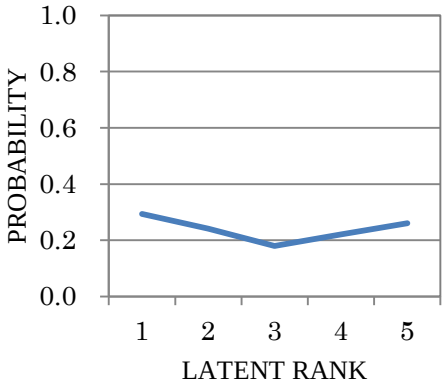

図 27　IRP の例 2

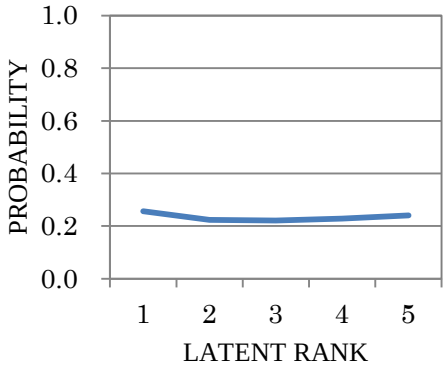

図 28　IRP の例 3

γ=1 でなくても（たとえば γ=0.5 でも），図 27 のように c の値が大きい場合（c=0.113）も，除去すべきであろう．（1）（2）の場合は，より上位のランクの受験者の正答確率の方が低い項目という意味で望ましくない．

c の値がある程度以下の値になると（0 に近づくと）あまり大きな問題ではなくなる．それよりも a の値の小ささの方が問題になる．つまり，この状況は，図 28 のように IRP がほぼ平らな状態の項目であることを示すからである．（3）の場合は，各ランクの受験者の正答確率が同じで，各ランクの受験者の能力を識別するという意味において，望ましくない項目といえる．

上の（1）～（3）の要素を持つ項目が，LRT-CAT のアイテムバンク構築において望ましくない項目であり除去すべ項目として，どのような基準と手順で，それらの項目を除去すべきであろうか．次の優先順位を持つ 3 つの条件に照らし合わせ，1 項目ずつ除去し，再分析を行うという指針を提案する．ある時点の分析結果で，複数の項目が望ましくない要素を持っていいたとしても，一度に複数の項目を除去することは危険であろう．なぜなら，分析から 1 つの項目を外すだけで，他の IRP が大きく変わることもあり得るからである．

第 1 条件：γ=1

第 2 条件：$c \geqq 0.05$

第 3 条件：$a < 0.05$

まず，第 1 条件に当てはまる項目を探し，複数見つかった場合は第 2 条件を加え，c の値が最大のものを除去する．第 1 条件にも第 2 条件にも当てはまる項目が無くなった場合，第 3 条件を加え，複数ある場合は a 値が最小のものから除去する．

多肢選択形式（4 択）の英語文法語彙問題 80 項目に対する 207 人の受験生の応答データを，Exametrika を使い，単純増加制約をつけずに，潜在ランク数 5 で，SOM のメカニズムによる LRT に基づく分析を行い，前述の指針により，1 項目ずつ望ましくない項目を除去し，再分析を繰り返した．13 項目を除去し 14 回目の分析をしたところで，前述の指針に当てはまる項目は 1 つもなくなった．除去された 13 項目の，各分析時点での IRP と IRP 指標を示すと表 4 のようになる．除去の判断として使われた IRP 指標の部分を分かるようにするために，そのセルを網掛けしてある．たとえば，最初の Vgm0047 という項目の場合，γ=1 となる項目が他にもあったため，c の値も参照して除去する項目を決めている．第 1 条件で除去されたのは 4 項目，第 2 条件と第 3 条件で除去されたのは，それぞれ 4 項目と 5 項目であった．

LRT による分析に使ったものと同じ項目応答データを使って，RM に基づく項目除去を WINSTEPS により行った．基準は，これまで行ってきた「$MNSQ$＞1.3」かつ「$Zstd$＞1.96」を使った．前節の LRT の場合と同様に，最も望ましくない項目から除去し，1 項目を除去するごとに，分析をやり直した．その結果，8 項目が除去され，9 回目の分析で，上記の条件に引っかかる項目は無くなった．表 4 の結果と比較するために，表 4 に現れたすべての項目を表 5 にも掲載し，RM のミスフィットの基準で除去された項目名とその基準に使った値の部分を網掛けにした．

表 4　LRT（IRP指標）に基づく望ましくない項目の除去 [8]

Item	IRP					IRP Index					
	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	α	a	β	b	γ	c
Vgm0047	0.390	0.308	0.243	0.206	0.168	1	0.000	1	0.390	1.000	0.221
Vgm0042	0.266	0.213	0.187	0.154	0.101	1	0.000	1	0.266	1.000	0.165
Vgm0038	0.294	0.241	0.180	0.221	0.261	3	0.041	1	0.294	0.500	0.113
Vgm0008	0.196	0.194	0.186	0.154	0.149	1	0.000	1	0.196	1.000	0.047
Vgm0028	0.253	0.248	0.243	0.215	0.208	1	0.000	1	0.253	1.000	0.045
Vgm0007	0.241	0.275	0.299	0.267	0.205	1	0.034	3	0.299	0.500	0.094
Vgm0061	0.308	0.360	0.318	0.297	0.376	4	0.079	5	0.376	0.500	0.063
Vgm0014	0.919	0.958	0.979	0.958	0.919	1	0.039	1	0.919	0.500	0.060
Vgm0065	0.120	0.125	0.110	0.115	0.115	3	0.005	2	0.125	0.500	0.015
Vgm0032	0.256	0.223	0.221	0.229	0.240	4	0.011	1	0.256	0.500	0.035
Vgm0071	0.151	0.154	0.144	0.163	0.167	3	0.020	5	0.167	0.250	0.011
Vgm0056	0.105	0.133	0.132	0.152	0.165	1	0.028	5	0.165	0.250	0.001
Vgm0053	0.232	0.271	0.282	0.263	0.266	1	0.039	3	0.282	0.250	0.020

表 4 と表 5 を比べると，RM に基づいて望ましくない項目として除去された 8 項目は，すべて本研究が提案する LRT-CAT のためのアイテムバンク構築において望ましくない項目の除去指針によっても除去されている. 表 4 で除去されたが表 5 で除去されていない項目についても，RM のミスフィット指標の数値を見ると，どれも除去する基準に近い値のものばかりである.

まだ 80 項目について検討を加えただけであるので，ここから結論を導くのは早急すぎる. しかし，本研究で提案した LRT-CAT のためのアイテムバンク構築において望ましくない項目の除去指針は，RM のミスフィットによる除去と重なる部分が多く，おおむね良好に機能しているように思われる.

両者で判断が異なる部分にどのような特徴があるかについては，まだ解明できていない．今後より多くのテスト項目を分析し，明らかにしていく必要がある．また，提案した 3 つの条件式の中の値（特に c の値と a の値）は，ヒューリスティックなものであり，今後の研究と目的に応じて変化させるべきものだと考える.

[8] Exametrika では，c の値にマイナスの符合がついて出力されるが，表 4 中では，「減少した大きさの和」という定義にしたがい，マイナスの符合を取り除いた.

表 5　RM（ミスフィット指標）に基づく望ましくない項目の除去

	Measure	*SE*	Infit *MNSQ*	Infit *Zstd*	Outfit *MNSQ*	Outfit *Zstd*
Vgm0047	1.28	0.17	1.25	3.06	1.49	3.86
Vgm0042	1.75	0.19	1.18	1.55	1.50	2.81
Vgm0038	1.28	0.16	1.21	2.65	1.48	3.84
Vgm0008	1.64	0.18	1.19	1.74	1.48	2.92
Vgm0028	1.49	0.17	1.18	1.92	1.47	3.19
Vgm0007	1.31	0.16	1.24	2.88	1.49	3.82
Vgm0061	0.92	0.15	1.14	2.31	1.24	2.60
Vgm0014	-2.53	0.25	1.00	0.04	1.49	1.60
Vgm0065	2.39	0.21	1.10	0.69	1.38	1.61
Vgm0032	1.44	0.17	1.16	1.82	1.49	3.45
Vgm0071	2.03	0.19	1.12	1.02	1.30	1.59
Vgm0056	2.28	0.21	1.08	0.62	1.34	1.52
Vgm0053	1.28	0.16	1.18	2.28	1.33	2.74

4.2.　LRT-CAT アルゴリズムの提案

LRT に基づく CAT（LRT-CAT）リズムについて，木村・永岡(2011a)は，項目選択について初めて提案初めて行ったものである．木村・永岡(2012a)は，それに終了条件についての考察を加え LRT-CAT アルゴリズム全体を提案し，シミュレーションと実テストによってその提案を検証したものである．ここにそのアルゴリズムを述べる．

(1) How to START

ほとんどの場合，テストを開始する時点で受験者の能力は未知である．IRTに基づくCAT（IRT-CAT）の場合, 初期能力推定値は能力推定値（θ）の平均（θ=0.0）とすることが多い. LRT-CATの場合は「1,…,5」のような潜在ランクを用いて評価するため，その中央値を利用する方法が考えられる．しかし，この方法は分析する潜在ランク数が奇数の場合はよいが，偶数の場合は1つの潜在ランクに定まらず，うまくいかない．LRT-CATの場合，初期能力推定値はRMPを使って，一様分布とする方法を提案する．すなわち，分析する潜在ランク数をQとした場合，各潜在ランクに所属する確率を$1/Q$とする．

先に述べたattenuation paradox (Lord & Novick, 1968)を回避するために, 最初の出題は1問ではなく，潜在ランクの中央値付近の問題を複数問出題するテストレット形式とすることを提案する．具体的には, 分析する潜在ランク数が奇数の場合は, 中央のランクを中心に3ないし5の奇数問を, 偶数の場合は中央の2つのランクを含む2ないし4問を出題することとする．分析する潜在ランク

数が5以下の場合は, アイテムバンクから各ランクの問題を1問ずつランダムに選び出題するのがよいだろう.

離散的なLRTでも，IRP指標βを利用して，様々な困難度の項目を組み合わせて項目の選定を行うことは実用上十分可能であり，IRTやRMよりもむしろ簡便である.

(2) How to CONTINUE

潜在ランクの中央値付近の複数問からなるテストレット（分析する潜在ランク数が5以下の場合は，各潜在ランクから1問ずつとする）に対する受験者i の応答ベクトルを$\mathbf{u}_i$とし，テストレットに選ばれた項目のIRP（行ベクトル$\boldsymbol{\upsilon}_j$）を集めた行列を$\mathbf{V}^{(0)}$，受験者が潜在ランクR_qに所属する確率をF_qとすると，次式により受験者iの暫定RMP（ベクトル$\mathbf{p}_i$）が求まる.

$$\mathbf{p}_i = \left\{ p_{iq} \middle| p_{iq} = P\left(F_q = 1 \middle| \mathbf{u}_i, \mathbf{V}^{(0)}\right) \right\} \qquad (Q \times 1)$$

$$P\left(F_q = 1 \middle| \mathbf{u}_i, \mathbf{V}^{(0)}\right) = \frac{p\left(\mathbf{u}_i \middle| \boldsymbol{\upsilon}_q^{(0)}\right)\pi_q}{\sum_{r=1}^{Q} p\left(\mathbf{u}_i \middle| \boldsymbol{\upsilon}_r^{(0)}\right)\pi_r} \tag{32}$$

ここでπ_qは, 潜在ランクR_qに所属する事前確率であり, 初期値は一様分布$\pi_q = 1/Q$（q=1,2,…,Q）とする.

この受験者i の暫定潜在ランクは，RMP（ベクトル$\mathbf{p}_i$）の中で，最も値の大きい潜在ランクであるから，IRP指標βがこの暫定潜在ランクと同じ項目をアイテムバンクから出題するという方法も考えられる．しかし，LRTがIRTと大きく異なるのは，前述したように，受験者の特性を，一義的に推定された潜在ランクとしてとらえるだけでなく，RMPとして多義的に表現できる点である．このLRTの特徴を生かすために，本研究ではRMPに基づき項目を選択する方法として，暫定RMPとIRP の差分ベクトルの積和平均による項目選択ルールを提案する.

いま，ランク数をQとし，受験者i がn項目解答した時点の暫定RMPを

$$\mathbf{p}_i^{(n)} = [\, p_{i1}^{(n)} \cdots p_{iQ}^{(n)} \,]' \qquad (Q \times 1) \tag{33}$$

とする．ここで，$p_{iq}^{(n)}$ は，受験者iのR_qに対する暫定的な所属確率である.

また，アイテムバンクにあるj 番目の項目のIRPを

$$\boldsymbol{\upsilon}_j = [\upsilon_{j1} \cdots \upsilon_{jQ}]' \qquad (Q \times 1) \tag{34}$$

とする．ここで，υ_{jq}はR_qに所属する受験者の項目j に対する正答確率である．さらに，IRPの差分ベクトル

$$\boldsymbol{\delta}_j = [\delta_{j1} \cdots \delta_{jQ-1}]' \quad ((Q-1)\times 1) \tag{35}$$

を計算する．ここで，

$$\delta_{jq} = \upsilon_{jq+1} - \upsilon_{jq} \quad (q = 1,2,\cdots,Q-1) \tag{36}$$

である．

そして，$\mathbf{p}_i^{(n)}$ に対する識別度の高さを以下の式を用いて評価する．すなわち，

$$\lambda_j^{(n)} = \frac{\sum_{q=1}^{Q-1} p_{iq}^{(n)} \delta_{jq} + \sum_{q=1}^{Q-1} p_{iq+1}^{(n)} \delta_{jq}}{2} \tag{37}$$

である．

このλの値が大きいほど受験者i に対する識別度が高いことを示す．これはIRTの項目選択ルールについて論じているVan der Linden （1998）の中に出てくるMaximum Expected Posterior Weighted Information という方法に相当すると考えられる．

識別度の高いもの，すなわち，λの値が最大のものから選択する方法も考えられるが，CAT初期で識別度の高いものから実施すると，アイテムバンクがあまり大きくない場合，CAT終期でRMPが収束し始めたときに，識別度の高いアイテム（局所的に（受験者の暫定ランクの付近で）急峻なIRPを持つ項目）がなく，かえって効率が悪くなることが懸念される．CAT初期の暫定RMPはなだらかな形状であると考えられるので，識別度の低いものを実施し，識別度の高いものをCAT終期に温存しておく方がよいのではないだろうか．本研究のLRT-CATの項目選択ルールとしては，アイテムバンクの中でλが最小のものから選択することを提案する．

アイテムバンクに蓄えられた項目数が多くなると，計算負荷がサーバーにかかり過ぎるので，本研究ではIRP指標βが受験者i の暫定の潜在ランクの推定値±１の項目に限定してλの値を計算し，その中で最小となる項目を選択することとした．

(3) How to STOP

IRT-CATの場合，θのSEが十分小さくなった場合に推定が収束したと判断して終了するか，あらかじめ決めておいた項目数（m）に達した場合に終了にするのが一般的である．LRT-CATの場

合, 項目数がm に達した場合に終了にする方法の他に, RMPの変化が一定以下になった場合に, RMPの推定が収束したと判断して終了させる方法が考えられるだろう. 一般的にCATの初期の段階ではRMPの変化は大きく，受験項目数が増えるにしたがって，その変化は小さくなると考えられるからである.

$\mathbf{p}_i^{(n)}$ と $\mathbf{p}_i^{(n+1)}$ の差ベクトルの要素のうち，絶対値の最大値

$$\mu_{jq}^{(n)} = \max_{q \in Q} | p_{jq}^{(n)} - p_{jq}^{(n+1)} | \tag{38}$$

の値を基準として，たとえば μ が0.05未満になったときにCATを終了させるという条件で実施することも可能である．終了条件をどのようにすべきかについては，実践編の8.1で，シミュレーション研究の結果を踏まえ再検討を行う.

II. 実践編

ここでは，大学に入学してくる学生の英語力を測定し，能力別クラス分けを行うための小規模な CAT を開発することを目標に行われた実践的研究について報告するとともに，開発過程で見つかった問題について考察を加える．理論編で述べた 5 段階の CAT 開発フレームワークにそって，第 1 段階から第 5 段階までででどのような実践的研究が行われたか整理すると次の表 6 のようになる．

表 6　CAT 開発のフレームワークにそって行われた実践的研究

CAT 開発のフレームワークの段階	実践的研究
第 1 段階： 実現可能性と適用性の評価， 計画調査の段階	・目的や利用規模からどのようなテスト理論とモデルを利用すべきかについての検討 ・利用可能なオープンソースの検討
第 2 段階： アイテムバンクの項目作成あるいは既存のバンクの活用段階	・どのような項目を使うかの検討 ・使おうとしている項目が目的に照らして妥当であるかの検討
第 3 段階： 事前テストの実施とアイテムバンクの項目特性を分析する段階	・事前テストの実施と項目分析（二値モデルと多値モデル） ・アイテムバンクの拡充と統合
第 4 段階： 最終的な CAT の仕様を決定する段階	・シミュレーションによる CAT 仕様の検討 ・アイテムバンクの検証
第 5 段階： 実際に CAT を世に出す段階	・CAT を実装するためのシステム開発

ただし，第 5 章以降で述べる実践的研究は，この 5 段階の順に進められたわけではない．開発途中に新しい知見が得られることや，新しいソフトウェアが発表されることはしばしばあり，そのたびごとに前の段階に戻り再検討を加えて，また先に進めて行かなければならない．つまり，実践的研究は開発のフレームワークの段階を行きつ戻りつしながら進められる．

第 3 段階の事前テストは 3 年計画で実施する計画を立てたので，第 1 次事前テストが終了した段階で，その項目分析結果を元に，項目を固定して実施するプレイスメントテストを作成した．その評価を行うことで，第 2 段階で計画した項目が目的に照らして妥当であるかを検証し，さらに，第 2 次・第 3 次事前テストを進めた．また，第 2 次・第 3 次事前テストと並行して，CAT を実装するためのシステム開発を行った．つまり，第 2 段階から第 4 段階の CAT 開発の実践は，時系列的には並行して進められた．

第 5 段階の「CAT を実装するためのシステム開発」については，理論編の第 3 章と第 4 章で論じているので実践編では改めて述べない．RM に基づいた CAT を実装するためシステムとし

ては，RM に基づいた CAT プログラム UCAT（Linacre, 1987）を元に開発した Moodle UCAT モジュール（Kimura, Ohnishi & Nagaoka, 2012）について，すでに理論編 3.3 で詳述している．LRT に基づいた CAT を実装するためのシステムとしては，すでに理論編 4.2 で木村・永岡(2011a, 2012a)が提案した CAT アルゴリズムに基づき，LRT-CAT モジュール（秋山・木村・荘島, 2011）が開発されている．

さらに，CAT 開発と並行して，受験者に対して行ったアンケート調査を元に CAT の心理学的側面について考察を加え，CAT を実装するためのシステム開発の参考にした．また，CAT の結果を何らかの CDS と結び付けられないかの検討を行うとともに，CAT の結果を受験者に提示する際に，診断的要素を加える方法についても検討を加えた．

5. CAT 開発フレームワーク第１段階での実践的研究

CAT 開発を始めるにあたって，どのテスト理論に基づいて進めるかは重要である．また，開発の各段階で利用可能なオープンソースとしてどのようなものがあるのかについても整理しておく必要がある．本研究の CAT 開発はどのテスト理論に基づいて行うべきなのかについては，すでに理論編 1.6 で論じた．ここでは，一般的に CAT 開発のためにどのようなオープンソースが利用できるかについて，筆者が企画した IACAT 2012 Conference のシンポジウム（Kimura, Han, Kosinski, & Shojima, 2012）での議論をもとに紹介する．

5.1. オープンソースとフリーウエアの検討

理論編で述べたCAT開発フレームワークの各段階では，第2段階（項目作成段階）を除いて，専用のソフトウェアが必要になる. 第1段階と第4段階ではシミュレーションのためのソフトウェア，第3段階では項目分析のためのソフトウェア，第5段階ではCATを実装するシステムが必要となる．

商用ソフトウェアであれば, データの生成とCATシミュレーションのためのソフトウェアとしてCATSim (Weiss & Guyer, 2010)，項目分析としてはRM用のWINSTEPS (Linacre, 2009) やRUMM (Andrich et al, 2010), IRT用のBILOG-MG (Zimowski et al, 2003) やMULTILOG (Thissen et al, 2003), CAT実装のためのソフトウェアとしてFastTEST (Assessment Systems Corporation & 4ROI, 2010) などがある．

本論文のテーマは，小規模なCATを前提としたオープンソースによるCAT開発であるので，本節ではCAT開発に利用可能なオープンソース（ソースの公開はされていないがフリーソフトウエアであるものを含む）を利用することにした．第6章以降の実践的研究は，3年間かけて行われたもので, 開発のフレームワークの段階を行きつ戻りつして進められたので，ここに紹介するものをすべて，第6章以降の実践的研究で使用したわけではない．また，ここではCAT開発に利用できるオープンソースを網羅的に紹介するのではなく, IACAT2012のシンポジウムA framework and approaches to develop an in-house CAT with freeware and open source software (Kimura et al, 2012) で

取り上げたものを中心に紹介することにする.

5.1.1. データの生成とシミュレーションのためのオープンソースとフリーウエア

オープンソースとしては，統計解析とグラフィックスを行うための言語であり環境 R のパッケージの一つである catR (Magis & Raîche, 2012) でデータの生成とシミュレーションができる.

RはGNUプロジェクト [9]の一つであり，ベル研究所でChambersらにより開発されたS言語・環境に似ている. 多様な統計手法（線形・非線形モデル，古典的統計検定，時系列解析，判別分析，クラスタリング，その他）とグラフィックスを提供し，広汎な拡張が可能である．Rに関する情報はすべてR Projectのホームページ [10]から入手することができる．Rについて日本語での情報交換を目的に作られたRjpWiki[11]もある.

catR は，R 環境において，4 パラメータ以下のロジスティックモデルで分析された既存のアイテムバンクまたは，パラメータの分布を指定することで生成したアイテムバンクを生成して，シミュレーションを行うことができる. いくつかの初期項目選択方法とそれ以降の項目選択方法を指定し，異なる能力推定法（maximum likelihood, Bayes modal, expected a posteriori, weighted likelihood）により推定を行い，3 通りの終了条件（指定項目数，推定の精度，受験者の弁別）で CAT のシミュレーションを行うことが可能である．分析結果を容易にグラフ形式で結果を出力させることも可能である.

ソースは公開されていないが，フリーウエアとしては，SimulCAT (Han, 2012) が多くの機能を備えており使いやすい．SimulCATもcatRと同様，既存のアイテムバンクあるいは，パラメータの分布を指定して発生させたデータを利用してシミュレーションを行うことができる．SimulCATでは，多様な項目選択ルールを扱える：広く使われている 6 種類（maximized Fisher information (MFI: Weiss, 1982), a-stratification (Chang & Ying, 1999; Chang, Qian, & Ying, 2001), global information (Chang & Ying, 1996), interval information, likelihood weighted information (Veerkamp & Berger, 1997), gradual maximum information ratio (GMIR: Han, 2009), efficiency balanced information (EBI: Han, 2010)）と，item exposure をコントロールした 4 種類（randomesque strategy (Kingsbury & Zara, 1989), Sympson and Hetter method (1985), multinomial methods—both conditional and unconditional (Stocking & Lewis, 1995, 1998), fade-away method (FAM: Han, 2009)），さらにコンテンツ・バランスについては，Kingsbury & Zara (1989)のcontent script method and the constrained CAT methodもサポートしている．SimulCATの特徴は，多様な項目選択ルールを扱えることだけでなく，わかりやすいグラフィカルないインターフェースにもある．ソフトウェアと

---

[9] Unix ライクなオペレーティング・システムをフリーソフトウエアとして開発するために、1984 年に発足したプロジェクトで、アプリケーション、ライブラリ、開発ツール、そしてカーネルと呼ばれるリソースを割り当てハードウェアとやりとりするプログラム、からなるソフトウェアのコレクションである（http://www.gnu.org/）.

[10] http://www.r-project.org/

[11] http://www.okada.jp.org/RWiki/

マニュアルはSimulCATのホームページ[12]からダウンロードできる.

5.1.2. 項目分析のためのオープンソースとフリーウエア

オープンソースとしては，シミュレーションソフトの catR と同様に R パッケージとして ltm（Rizopoulos, 2006）がある．RM と 2PLM，3PLM など IRT の項目分析ツールだが，CTT の範疇の IT 相関やコロンバックのアルファ係数なども計算できる．2 つの等化の手法（alternate form equating, across sample equating），多様なグラフ出力（ICC，IIC，TIF，SEM，item person map など），モデルの適合度指標（RM 用に bootstrap Pearson χ^2, 2PLM と 3PLM 用に AIC や BIC など），item-fit ならびに person-fit を判断する統計量などを求めることができる．2 値データだけでなく多値データの分析も graded response model と generalized partial credit model によって可能である．

フリーウエアは数多く存在するが，IRT と LRT のモデルの両方を扱える Exametrika (Shojima, 2010) を紹介する．Exametrika は，IRT の二値モデル（dichotomous model），ボックの名義モデル（Bock’s nominal model），サメジマの多値モデル（Samejima’s graded model）を扱うことができ，パラメータは 2 の場合から 5 の場合まで指定できる．出力オプションとして，適合指標と IRF のグラフも出力することもできる. また, 固定項目を指定して等化を行うことができる. さらに，LRT 二値モデル（dichotomous model），名義モデル（nominal model），多値モデル（graded model）を扱うことができ，事前分布を指定することや目標潜在ランク分布（一様分布または正規分布）を指定することや，短調増加制約をつけて分析することもできる．推定方法については，GTM と SOM の２つが用意されている．出力のオプションとして，適合指標や IRP のグラフを出力することもできる．また，固定項目を指定して等化を行うことができる．IRT と LRT のモデル以外にも，非対称三角尺度法（asymmetric triangulation scaling, ATRISCAL: Shojima, 2012）による分析や，カテゴリカルデータ解析（categorical data analysis , CDA)の分析機能も用意されている．ATRISCAL は非対称多次元尺度法の 1 つであり，項目間のグローバルな従属関係を記述する多変量解析モデルである．Exametrika の CDA では，閾値，平均情報量（entropy），項目得点双列相関（biserial correlation coefficient）・項目得点多列相関（polyserial correlation coefficient），項目間四分相関（tetrachoric correlation coefficient）・項目間多分相関（polychoric correlation coefficient）を出力する．

Exametrika の特徴は，1 つのソフトウェア上で，多様なモデルの中から適切なものを選択して分析が可能なことと, インターフェースがわかりやすく, Excel のシートからデータを読み込み，分析結果を Excel の別シートに出力して保存できることである．

Exametrika が発表されるまでは，IRT の分析には EasyEstimation シリーズ（熊谷, 2009）を，LRT の分析には neutet(Hashimoto & Shojima, 2007)を利用した．いずれも，ソースは公開されていないが，プログラムが WEB 上に公開されたフリーウエアである．また，RM の分析には，当初書籍に添付されたプログラム TDAP(Ohtomo et al, 2002)を利用していたが，より細かな分析

[12] http://www.hantest.net/simulcat

ができる有償プログラム WINSTEPS(Linacre, 2009)に変更した（WINSTEPS には機能を制限した無償版プログラム MINISTEP がある）.

5.1.3. CAT を実装するためのオープンソース

オープンソースでCAT実装する方法として，ここでは２つのアプローチについて述べる．ひとつは，オープンソースのLMSであるMoodle[13]の上に追加モジュールを開発してCATを実装する方法である．もうひとつは，Cambridge University Psychometrics Centerが，開発しオープンソースとして公開しているCAT実装のためのプログラムConcerto[14]を利用する方法である．

Moodle は多様な機能をもつ LMS であり，日本を含め世界中の多くの教育機関や企業等で教育に利用されている．いろいろな形式で質問を作り出題し，採点・管理する機能を持っている．CAT を実装する機能はないが，オープンソースであるので，CAT を実装するモジュールを開発し，組み込むことが可能である．現在のところ，RM-CAT としては，理論編 3.2 で詳しく説明した UCAT（Linacre, 1987）を元に開発した Moodle UCAT モジュール（Kimura, Ohnishi & Nagaoka, 2012）が, LRT-CAT としては, 理論編 4.2 で提案した LRT-CAT アルゴリズム（木村・永岡, 2011a）に基づき開発された LRT-CAT モジュール（秋山・木村・荘島，2011）がある．

Concertoは 2011 年 7 月に初めて公開された複合的なプログラムであり, HTMLの表現の柔軟さと，R環境の強力な計算能力と，MySQLの安全なデータベース機能を組み合わせて開発されたものである．Concertoのプログラムと情報はConcerto Projectのホームページ [15]からすべて入手することができる．ホスティング・サービスも提供されているので，利用者が手元にサーバーを構築しなくても，すぐにConcertoを利用して，CATを実装する環境を手に入れることができる．1 ヶ月に 150 までの応答者数なら，無料ですべての機能が利用可能で，かつメールによるサポートを受けられるホスティング・サービスもある．

次章以降の実践的研究では，Moodle 上に追加モジュールを開発して CAT を実装するアプローチをとったが，Moodle の基幹プログラムがバージョンアップされるたびに，開発したモジュールを修正する必要があるので，注意が必要である．一方，Concerto にはそのような問題は発生しないが，LMS としての機能は基本的に備えていない．今後は，オープンソースの CAT 実装プログラムである Concerto と，オープンソース LMS である Moodle の間でデータ連携を図るアプローチが有効であると考える．

6. CAT 開発フレームワーク第２段階での実践的研究

CAT 開発フレームワーク第２段階での実践的研究については，本研究で使用した項目がどのようなものであったかと，その項目についての妥当性を検討した結果について報告する．

---

[13] http://moodle.org/
[14] http://www.psychometrics.cam.ac.uk/
[15] http://code.google.com/p/concerto-platform/

6.1. CAT のために用意する項目について

本研究で利用した項目は，すべて日本英語検定協会の許可を得て，英検準 1 級から 3 級の過去問題（2007～2008 年度）を利用した．項目の形式は，すべて 4 択の多肢選択問題で，項目の種類としては，文法語彙問題（vocabulary and grammar, Vgm），ダイアローグの聴解問題（listening comprehension with dialogues, Dlg），モノローグの聴解問題（listening comprehension with monologues, Mlg），読解問題（reading comprehension, Rdg）の 4 種類である．それぞれの種類の項目の例は，図 29～図 32 に示すとおりである．当初それぞれ 4 つの独立したアイテムバンクとして構築していたが，のちに Dlg と Mlg はアイテムバンクを統合して 1 つの聴解問題（listening comprehension, Lng）のアイテムバンクとした．Rdg だけは 1 つのパッセージに対して 2～5 問の質問がある形式なので，多値型モデルとして分析された（その他は 2 値型モデル），いずれのアイテムバンクについても，同じデータセットに対して RM と LRT の 2 通りの分析が行われた．

(1) My school is looking for ( ) to clean the park. I'm going to go this Sunday.

1 villages **2** customs **3** cultures **4** volunteers

(2) ***A*** *:* How do I get to the library?

B *:* ( ) the bridge and go straight. It's on the right.

1 Cross **2** Put **3** Break **4** Lend

図 29　Vgmの項目例 [16]

スクリプト	選択肢
☆☆No. 11 ☆Let's go to the museum this weekend, Bill. ★Sorry, Asako. I'm going to Kyoto with some friends from America. ☆OK, I see. Well, have a good trip. ★Thanks. ☆☆**Question:** Where will Bill go this weekend?	*No. 11* **1** To America. **2** To Kyoto. **3** To a museum. **4** To Asako's house.

図 30　Dlg の項目例 [16]

[16]英検 3 級 2008 年度 第 1 回より引用

スクリプト	選択肢
☆☆No. 21 ☆Next month, Mary and her older sister Erin will visit Japan. They'll stay in Osaka. It'll be Mary's first time in Japan, but Erin has visited three times before. They're looking forward to it. ☆☆Question: How many times has Erin visited Japan?	***No. 21*** **1** Never. **2** Once. **3** Twice. **4** Three times.

図 31　Mlg の項目例 [16]

本文	項目
HAVE YOU SEEN MY RACKET? I can't find my tennis racket. I need my racket for the Greenmount City High School Tennis Tournament on June 28th! I left it at the bus stop in front of the park on Wednesday, June 12th. Please help me! **Description:** My racket is blue and red, with a brown handle. The brand name is Swift. The racket is in a black bag. My name is on the inside of the bag and on the racket. If you know where my racket is, please contact me at 333-4567. I'm at home after 5 p.m. My name is Caroline Jimenez.	Q1. Where did Caroline lose her tennis racket? 1. At the school gym. 2. At the bus stop. 3. At a tennis tournament. 4. At a sports store in Greenmount City. Q2. lf people have seen Caroline's racket, they should 1. call her after 5 p.m. 2. meet her in the park. 3. bring it to the tennis tournament. 4. give it to Greenmount High School.

図 32　Rdg の項目例 [16]

6.2.　用意した項目の妥当性の検討

後述のCAT開発の第3段階の一部である第1次事前テストを分析した結果から, 項目を固定して行うプレイスメントテストが作成された．第3段階の事前テストをさらに進める前に，このプレイスメントテストを評価することで, 準備を進めているCATで使用する予定の項目で, 意図して

いるとおり基礎的な英語力を測定できるかどうか検証することにした．そのために，以下に述べる方法で，このプレイスメントテストのスコアと，CASEC[17] とTOEIC Bridge[18]のスコアの相関分析を行った．

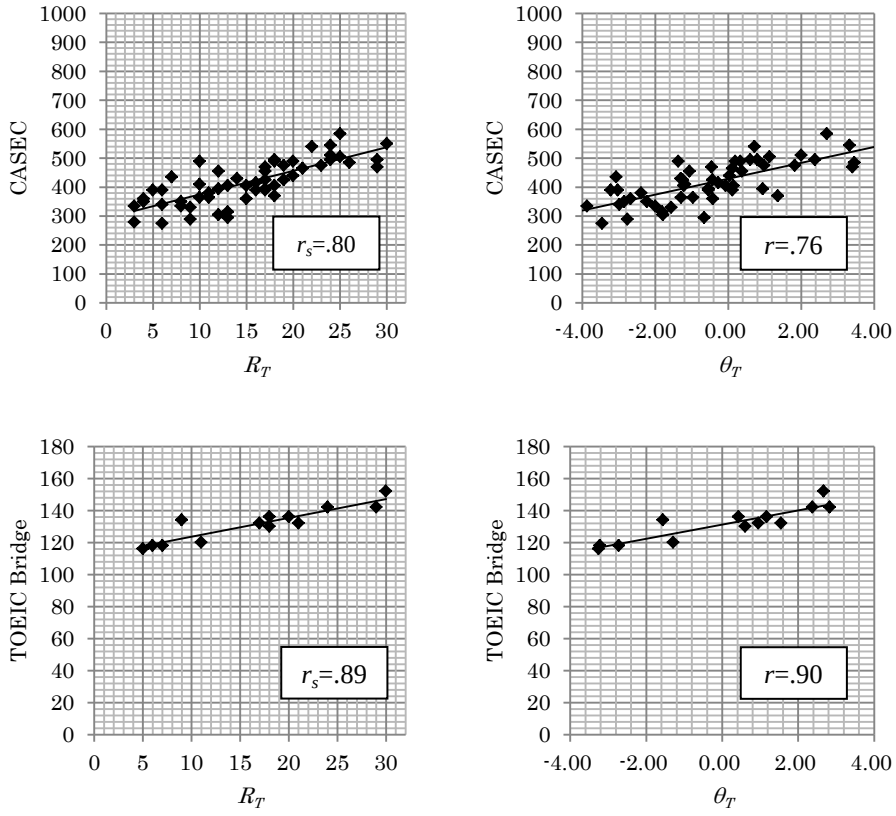

図 33　R_T，θ_Tと他の英語能力試験との相関

注：R_Tと他の英語能力試験との相関はスピアマンの順位相関係数 (r_s) を，θ_Tと他の英語能力試験との相関はピアソンの積率相関係数 (r) を用いた．

第1次事前テスト受験者の中で，プレイスメントテストに使われる項目すべてに解答している者が75人いた．このうち55人が，第1次事前テスト実施数週間後に行われたCASECを，13人がTOEIC Bridgeを受験していた．プレイスメントテストのスコアは，後述の7.3のクラス分けのシミュレーションで利用したスコアR_Tとθ_Tとし，それぞれCASECとTOEIC Bridgeの相関係数を計算した．その結果，LRTによって推定された潜在ランク (R_T) も，1PLMによって推定された潜在能力値 (θ_T) も，2つの英語能力試験の総合スコアと高い相関があり，特にTOEIC Bridgeとの相関は.89-.90と高かった（図33参照）．CASECとの相関よりもTOEIC Bridgeとの相関の方が高かった

[17] （財）日本英語検定協会が基礎開発し，現在（株）教育測定研究所が開発・運営しているインターネット上で受験できる英語コミュニケーション能力を評価する IRT に基づいた CAT．4 つのセクションからなり，各セクション 250，合計 1000 点のスコアが示される．

[18] ETS により基礎的なコミュニケーション英語能力を評価するために開発された世界共通のテストである TOEIC の特長を備えつつ初・中級レベルの英語能力測定に照準を合わせて設計されたテスト．リスニングセクション 50 問，リーディングセクション 50 問からなり，得点はそれぞれ 10 点～90 点の 2 点刻みで示される．

のは，CASECが想定している測定範囲よりも，TOEIC Bridgeが想定している測定範囲の方が，今回使用したテストにより測定している範囲が近いためだと推察される. サンプルサイズが大きくはないために，相関係数の推定値の精度（標準誤差）は必ずしも十分ではないが，高い点推定値が得られていると考えられる．この結果から，CATを実装するために準備を進めているアイテムバンクの項目は，おおむね妥当なものであると考え，予定どおり第2次・第3次事前テストを進めることとした.

7. CAT 開発フレームワーク第３段階での実践研究

事前テストは, 協力が得られた2大学の1年生の英語授業内の課題と授業時間外の自由課題として複数回に分けて実施された．各事前テストは項目の種類ごとに別テストとし，30分程度で解答できるようにするため，項目数は多くとも30項目程度とした．また，分析の信頼係数を確保するため，各項目について100～200人の受験者からの解答を集めることとした．CATを実施する各アイテムバンクに十分な項目を用意するため，事前テストは2008年から2010年にかけて，第1次から第3次に分けて実施・分析する計画とした.

最終的にCATを開発することを念頭に置きつつも，第1次事前テストの項目分析の結果から適切な項目を選び出し，固定された項目からなるプレイスメントテストを作成し(木村，2008c), 実際に利用した. そのプレイスメントテストに選ばれた項目をアンカー項目として，複数のテストを作成し, 第2次事前テストを実施した. 第2次事前テストまでの分析を終えた段階で, 教員間・学校間でアイテムバンクを共有することを計画し，アイテムバンクを公開した．第2次事前テストまでの研究成果について学会で発表し，そこで受けた質問やコメントを参考に，ミスフィット基準の見直しを行った. 新しいミスフィットの基準により，第1次から第2次事前テストまでのデータの再分析を行った上で，使用するアンカー項目を選び直し，第3次事前テストを実施した.

各事前テストの受験者の数をまとめると，表7～表10のようになる（Rdg以外はテスト単位で項目数や受験者数を示し，大問形式のRdgは大問ごとに項目数や受験者数を示した）．同一の受験者が複数のテストを受験した場合，前のテスト受験時期から3ヶ月以上経過している場合は，別受験者として扱い分析を行った．また，項目分析は項目の種類（Vgm，Dlg，Mlg，Rdg）ごとに行った．第1次事前テストでは，Vgm，Dlg，Mlgの3種類であったが，第2次・第3次事前テストではRdgが加えられた．また，2種類の聴解問題DlgとMlgは1つのアイテムバンクLngに統合された.

CAT 開発フレームワーク第３段階で行われた実践的研究として，以下に次のものを報告し考察を加える.

（1） ２値モデルの分析：Vgm, Dlg, Mlg の項目分析
（2） 固定された項目によるプレイスメントテストの作成
（3） 固定された項目によるプレイスメントテストによるクラス分けのシミュレーション
（4） 固定された項目によるプレイスメントテストの実施
（5） 多値モデルの分析：Rdg の項目分析

（6） アイテムバンクの拡充と項目困難度の等化

（7） RM におけるミスフィットの基準見直しと再分析

（8） RM におけるアイテムバンクの統合

表 7　Vgm の各事前テスト項目数と受験者数

	テスト名	項目数	うちアンカー項目	ミスフィット項目数	受験者数	合計
第 1 次事前テスト	2008A-Vgm	20		1	222	222
	2008B-Vgm	20				
	2008C-Vgm	20				
	2008D-Vgm	20				
第 2 次事前テスト	2009A-Vgm	32	16	3	292	1070
	2009B-Vgm	32	16		258	
	2009C-Vgm	32	16		268	
	2009D-Vgm	32	16		252	
第 3 次事前テスト	2010A-Vgm	26	6	2	284	1575
	2010B-Vgm	26	6		256	
	2010C-Vgm	26	6		224	
	2010D-Vgm	26	6		304	
	2010E-Vgm	26	6		295	
	2010F-Vgm	26	6		212	
合計受験協力者数						2867

表 8　Dlg の各事前テスト項目数と受験者数

	テスト名	項目数	うちアンカー項目	ミスフィット項目数	受験者数	合計
第 1 次事前テスト	2008A-Dlg	11		1	157	222
	2008B-Dlg	12				
	2008C-Dlg	12				
	2008D-Dlg	12				
第 2 次事前テスト	2009A-Dlg	16	7	1	297	1145
	2009B-Dlg	16	6		275	
	2009C-Dlg	16	6		283	
	2009D-Dlg	16	7		290	
第 3 次事前テスト	2010A-Dlg	29	6	0	263	1036
	2010B-Dlg	29	6		321	
	2010C-Dlg	30	6		310	
	2010D-Dlg	30	6		142	
合計受験協力者数						2403

表 9　Mlg の各事前テスト項目数と受験者数

	テスト名	項目数	うちアンカー項目	ミスフィット項目数	受験者数	合計
第 1 次事前テスト	2008A-Mlg	9		0	119	222
	2008B-Mlg	9				
	2008C-Mlg	9				
	2008D-Mlg	8				
第 2 次事前テスト	2009A-Mlg	16	9	1	277	1119
	2009B-Mlg	16	10		274	
	2009C-Mlg	16	10		282	
	2009D-Mlg	16	9		286	
第 3 次事前テスト	2010A-Mlg	24	6	0	198	763
	2010B-Mlg	24	6		257	
	2010C-Mlg	23	6		138	
	2010D-Mlg	23	6		170	
合計受験協力者数						2104

表 10　Rdgの各事前テスト項目数と受験者数 [19]

			第 1 次事前テスト					第 2 次事前テスト			第 3 次事前テスト			
大問名	英検級	小問数	G1	G2	G3	G4	合計	G5	G6	合計	G7	G8	合計	総合計
E-1	3 級	2	130				130				198	76	274	404
E-2		3	130				130				191	76	267	397
E-3		5	127				127				185	74	259	386
E-4/F-4	準2級	3	125	72	64		261	205	75	280	128	50	178	719
E-5/F-5		4	118	72	64		254	201	75	276	125	47	172	702
F-6	2 級	3		72	64		136				125	47	172	308
F-7		4		68	64		132				125	47	172	304
F-8		5		66	63		129				167	74	241	370
G-1	3 級	2	121		66		187							187
G-2		3	119		66		185							185
G-3		5	111		66		177							177
G-4/H-4	準2級	3	104	71	62	127	364	195		195				559
G-5/H-5		4	88	70	56	146	360	191		191				551
H-6	2 級	3		65		147	212	193	78	271				483
H-7		4		58		142	200	191	75	266				466
H-8		5		54		126	180	175	76	251				431
I-1	3 級	2						193		193				193
I-2		3						196		196				196
I-3		5						196		196				196
I-4/J-4	準2級	3						172		172	92	86	178	350
I-5/J-5		4						177		177	83	87	170	347
J-6	2 級	3						202		202	77	85	162	364
J-7		4						200		200	75	82	157	357
J-8		5						192		192	76	82	158	350
O-1	準1級	3									187	74	261	261
O-2		3									153	68	221	221
O-3		4									144	64	208	208
K-1	3 級	2							79	79	172	74	246	325
K-2		3							78	78	164	72	236	314
K-3		5							72	72	173	73	246	318
K-4/L-4	準2級	3							74	74	163	71	234	308
K-5/L-5		4							70	70	179	72	251	321
L-6	2 級	3							75	75	163	72	235	310
L-7		4							77	77	170	71	241	318
L-8		5							76	76	167	70	237	313
P-1	準1級	3									170	70	240	240
P-2		3									170	71	241	241
P-3		4									166	71	237	237
M-1	3 級	2									165	69	234	234
M-2		3									159	65	224	224
M-3		5									171	66	237	237
M-4/N-4	準2級	3									151	67	218	218
M-5/N-5		4									178	72	250	250
N-6	2 級	3							79	79	166	69	235	314
N-7		4							79	79	121	46	167	246
N-8		5							79	79	123	47	170	249
Q-1	準1級	3									78	87	165	165
Q-2		3									77	82	159	159
Q-3		4									78	66	144	144
合計		175	1173	668	635	688	3164	2879	1217	4096	5455	2642	8097	15357

7.1.　２値モデルの分析：Vgm, Dlg, Mlg の項目分析

第 1 次の事前テストは，表 4 に示すように Vgm，Dlg，Mlg それぞれ 4 つずつ 12 のテストレ

[19] 表中で網掛けになっている部分は，第 1 次事前テストの分析結果をもとに，第 2 次事前テストと第 3 次事前テストの分析の際にアンカー項目としたものである．

ットを作成し，2 つの大学の 1 年生合計 268 人に対して，Moodle の小テストモジュールを使って英語の授業時間内外で自由課題（実力問題）として実施した. テストレット全体で Vgm は 80, Dlg は 47，Mlg は 35 項目であった．テストレットの難易度がほぼ同じになるように，各テストレットには，英検の級からほぼ同数ずつが割り当てられるように構成した（表 11 参照）．データは Moodle から Excel 形式のファイルで取り出し，分析を行った．Moodle では，項目応答だけでなく解答開始時刻と終了時刻も記録されるので，最後まで解答しなかった者や，解答時間が異常に長い（あるいは短い）者もいたので，それらの解答を分析の対象から外し，222 人の受験者データを分析した.

表 11　各テストレットの項目の種類と数

Testlet	項目数	英検準1級	英検 2 級	英検準2級	英検 3 級
2008A-Vgm	20	7	5	5	3
2008A-Dlg	11	3	3	2	3
2008A-Mlg	9	--	4	3	2
2008B-Vgm	20	6	5	5	4
2008B-Dlg	12	3	4	3	2
2008B-Mlg	9	--	4	2	3
2008C-Vgm	20	6	5	5	4
2008C-Dlg	12	3	4	2	3
2008C-Mlg	9	--	4	3	2
2008D-Vgm	20	6	5	5	4
2008C-Dlg	12	3	4	3	2
2008D-Mlg	8	--	3	2	3

RM による分析は当初 TDAP (Ohtomo et al, 2002) を使って行われたが，その後アイテムバンクの拡張を行う段階で，詳細な分析を行うとともに，多くの等化作業を効率良く行うために，WINSTEPS (Linacre, 2009) を使って再度分析された. 項目についても受験者についても，*Zstd* の値が-2.0～＋2.0 の範囲を外れたものをミスフィットと判断して，分析対象から除外して再分析を繰り返した．最終的にそれぞれのアイテムバンクには Vgm に 36，Dlg に 13，Mlg に 19 項目が残った．このように多くの項目が削除されたことは，比較的受験者数が多いにもかかわらず，ミスフィットの判断に *Zstd* の値だけを使っていたことに原因があり，のちにミスフィットの判断を見直し，再分析を行うことになる（0 参照）．たとえば，Vgm については，ミスフィットの削除と再分析を 5 回行い，80 項目から半数以上の 44 項目が削除されたが，信頼性係数 KR-20 が 0.86 から 0.87 にしか上昇しなかったことからも，あまり効果的な項目削除は行われていなかったことが分かる．また，英検準 1 級の問題は，対象とした受験者集団には難しすぎ，正答率が低い上に多肢選択問題であるため偶然の正解も多かったようである．受験者の aberrant response の影響（Reise & Due, 1991）は大きい．確信度テスト法（張, 2007）の手法を利用することも有

効な手段かもしれない．後述の再分析では，こういった偶然の正解（あるいはうっかりミス）の扱いに関する工夫も行う. ミスフィット項目削除後の各アイテムバンクの困難度の統計情報は表 12 に示すとおりである．表 12 中の数値は logit を単位とするものだが，各アイテムバンクの項目の困難度がどの程度かを正答率によって, その基本統計量を整理すると表 13 のようになる（項目の種類ごとだけでなく，英検の級ごとに細分化して，正答率について基本統計量を示した）.

表 12　各アイテムバンクの困難度と SE の基本統計量（RM による分析：logit 単位）

Item Bank		*M*	*SD*	*Max*	*Min*
Vgm (n = 36)	Item difficulty (θ)	-0.707	1.097	1.609	-2.794
	SE of θ	0.160	0.034	0.267	0.135
Dlg (n = 13)	Item difficulty (θ)	-0.644	1.202	0.743	-2.782
	SE of θ	0.180	0.050	0.313	0.147
Mlg (n = 19)	Item difficulty (θ)	-0.455	0.875	1.221	-1.931
	SE of θ	0.188	0.024	0.244	0.171

表 13　各アイテムバンクの項目正答率についての基本統計量

種類	統計量	全体	英検準 1 級	英検 2 級	英検準 2 級	英検 3 級
Vgm	*N*	36	2	10	14	10
	M	64%	30%	51%	65%	80%
	SD	21%	16%	15%	18%	14%
	Max	94%	42%	66%	88%	94%
	Min	18%	19%	18%	29%	52%
Dlg	*N*	13	0	7	2	4
	M	63%	---	45%	82%	84%
	SD	23%	---	13%	9%	11%
	Max	95%	---	66%	91%	95%
	Min	34%	---	34%	73%	70%
Mlg	*N*	19	---	7	5	7
	M	60%	---	51%	52%	74%
	SD	18%	---	12%	18%	15%
	Max	88%	---	64%	73%	88%
	Min	24%	---	28%	24%	50%

理論的には LRT の枠組みでミスフィット統計量を計算することも可能であり，項目に関する適合度指標もいくつか開発されているが，第 1 次事前テストのデータ分析時（2008 年）にはま

だその計算を行うプログラムがなかったため，CTT の枠組みと，IRT の枠組みの両方の指標を利用して，ミスフィットの除去を行った．結果は上述の RM の場合の分析と同じであった．ミスフィットの除去を行った後のデータについて，LRT による分析を neutet（Hashimoto & Shojima, 2007）を使用して分析した．分析する潜在ランク数(Q)は 5 に設定した場合と 10 に設定した場合の 2 通りで行った．この段階での各アイテムバンクの項目の困難度を IRP 指標 β で整理すると表 14 と表 15 のようになる．

表 14　各アイテムバンクの困難度（LRT による分析，ランク数 5 の場合）

	IRP 指標 β					
Item Bank	1	2	3	4	5	合計
Vgm	14	6	7	3	6	36
Dlg	5	2	1	2	3	13
Mlg	5	3	5	3	3	19

表 15　各アイテムバンクの困難度（LRT による分析，ランク数 10 の場合）

	IRP 指標 β										
Item Bank	1	2	3	4	5	6	7	8	9	10	合計
Vgm	13	1	2	4	4	2	2	5	0	3	36
Dlg	4	1	1	1	1	1	0	2	0	2	13
Mlg	4	1	2	3	3	1	2	1	1	1	19

分析する潜在ランク数を 10 にすると，項目数が 0 や 1 になるランクが多く見られ，TRP では単調増加が確認されたが，IRP では単調増加を示さない項目がいくつか見られた．ランク数を 5 にすると，TRP だけでなく，すべての項目の IRP で単調増加が確認された．そのため，これ以降の分析においては，分析のランク数を 5 とすることにした．分析のランク数をいくつにすべきは，モデル適合度を示す指標を使って判断することが一般的だが，この段階ではまだ LRT のモデル適合度指標を計算するプログラムがなかったため，分析する項目数や受験者数，単調増加になる項目の割合から判断した．

7.2.　固定された項目によるプレイスメントテスト

LRT による分析を終えた段階で，固定された項目からなるプレイスメントテストを作成した（木村，2008c; 2009d）．その際，IRP を見て，いずれのランクにおいても 80%以上の正答確率で平板な 4 項目（すべて Vgm の項目）は，その解答結果から得られる情報が他の項目と比べて少ないので除外し，固定された項目からなるプレイスメントテストは，文法語彙問題の Vgm が 32 問，リスニング問題が Dlg の 13，Mlg の 19 を合わせて 32 問，総計 64 問のテストとした．

文法語彙問題とリスニング問題の割合が 50%ずつであり，予備テストの実施状況から考えて，分量的にもすべてを 45 分程度で実施できる実用的なテストとなった．

プレイスメントテストの各項目の困難度を整理すると，

表 16 のようになる．利用した項目が英検何級の問題であるかを示した上で， LRT による困難度（分析するランク数 (*Q*) を 5 にした場合の IRP 指標 *β*）と，RM による困難度 (*θ*) を示した．

表 16　プレイスメントテストの各項目困難度

・文法語彙問題(Vgm)			
項目番号	英検級	LRT (β)	RM (θ)
Vgm01	2 級	3	-0.17
Vgm02	2 級	3	-0.53
Vgm03	準 1 級	5	0.37
Vgm04	2 級	3	-0.19
Vgm05	2 級	2	-0.72
Vgm06	準 2 級	1	-1.34
Vgm07	2 級	5	0.55
Vgm08	準 2 級	3	-0.24
Vgm09	2 級	4	0.14
Vgm10	準 2 級	1	-1.92
Vgm11	3 級	1	-0.87
Vgm12	準 2 級	1	-0.92
Vgm13	3 級	1	-1.38
Vgm14	2 級	2	-0.63
Vgm15	準 2 級	2	-0.95
Vgm16	2 級	5	1.61
Vgm17	2 級	4	0.23
Vgm18	準 2 級	5	0.95
Vgm19	準 2 級	5	0.39
Vgm20	2 級	2	-0.58
Vgm21	準 2 級	2	-0.9
Vgm22	3 級	2	-0.92
Vgm23	3 級	3	-0.08
Vgm24	準 1 級	5	1.54
Vgm25	準 2 級	1	-1.17
Vgm26	3 級	1	-2.15
Vgm27	準 2 級	4	0.01
Vgm28	3 級	1	-1.47
Vgm29	準 2 級	3	-0.35
Vgm30	準 2 級	3	-0.08
Vgm31	3 級	1	-1.75
Vgm32	準 2 級	1	-1.87

・リスニング問題(Dlg・Mlg)			
項目番号	英検級	LRT (β)	RM (θ)
Dlg01	2 級	4	0.32
Dlg02	準 2 級	1	-2.19
Dlg03	2 級	2	-0.62
Dlg04	2 級	5	0.74
Dlg05	2 級	5	0.56
Dlg06	3 級	1	-0.78
Dlg07	3 級	1	-1.58
Dlg08	3 級	1	-1.97
Dlg09	2 級	3	-0.43
Dlg10	2 級	5	0.74
Dlg11	3 級	1	-2.78
Dlg12	準 2 級	2	-0.92
Dlg13	2 級	4	0.53
Mlg01	3 級	2	-0.94
Mlg02	準 2 級	1	-1.03
Mlg03	2 級	3	-0.31
Mlg04	2 級	2	-0.57
Mlg05	2 級	3	0.18
Mlg06	3 級	1	-1.66
Mlg07	準 2 級	3	-0.34
Mlg08	2 級	3	-0.16
Mlg09	3 級	4	0.03
Mlg10	3 級	1	-1.93
Mlg11	準 2 級	5	1.22
Mlg12	2 級	5	0.18
Mlg13	2 級	5	0.98
Mlg14	2 級	2	-0.50
Mlg15	3 級	1	-1.54
Mlg16	3 級	3	-0.34
Mlg17	3 級	1	-1.86
Mlg18	準 2 級	4	-0.16
Mlg19	準 2 級	4	0.10

RMによる分析とLRTによる分析がどの程度一致しているかを調べるために，アイテムバンクごとに両者のスピアマンの順位相関係数を求めた. Vgmで.97, Dlgで.91, Mlgで.89と非常に高く，項目困難度について両者は，ほぼ同様の推定が行われていることが確かめられた．図34はこれを図示したものである（木村，2009a）.

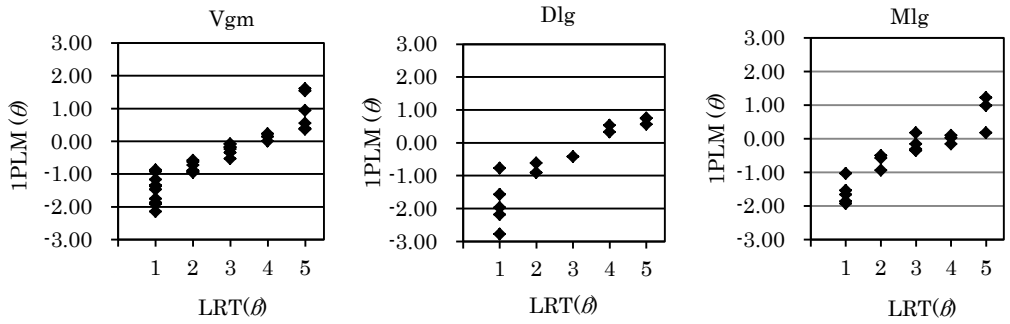

図 34　LRT の項目困難度 (β) と 1PLM の項目困難度 (θ) の比較

7.3.　クラス分けのシミュレーション

表 16 に示した 64 項目を使って実際にプレイスメントテストを実施する前に, 実施結果を LRT で分析した場合にどのような結果が得られか確認するために，事後シミュレーション（事前テスト受験集団の中で, この 64 問すべてを解答していた学生 75 人のデータを分析すること）により，クラス分けのシミュレーションを行った．つまり，各項目の IRP の値を事前テストで推定された値に固定した上で，改めて 75 人の解答パタンから，各人の潜在ランクの推定を行い，その推定結果に基づいてクラス分けを行った．

まず 3 つの問題の種類ごとに分析するランク数を 10 として分析を行った．問題の種類ごとに推定された潜在ランク（R_{vgm}, R_{dlg}, R_{mlg}）を多値の順序データとして扱い，LRT の拡張モデルの一つである段階ニューラルテスト（graded neural test, GNT）モデル（荘島, 2008b）によって，総合評価を求めるべきであるが, GNT モデルに基づいた計算を行うプログラムがまだなかったため, 全体的な傾向を判断することを主眼に，総合評価は 3 つのランク（R_{vg}, R_{dlg}, R_{mlg}）の和（R_T）を使うという簡易方法を用いた．各潜在ランクが 10 で分析されているので，R_T は 3〜30 となる．LRT による分析の妥当性を検証するために，同じデータについて，1PLM による能力推定も LRT の場合と同様の手順で行い，その結果を比較した．

R_Tを英語基礎力総合評価と，これに基づいて75人を，各クラスがだいたい同じ人数になるように分けると，図35のように14〜16人で5クラスに分けることができた．LRTによって分けた各クラスの英語力がどのような状況か，R_Tと 1 PLMによる3つの推定値（θ_{vgm}, $\theta_{dlg,}$ θ_{mlg}の和（θ_T）と，正答数S_Tをたクラスごとに比較したものが，表17である（木村，2009a）.

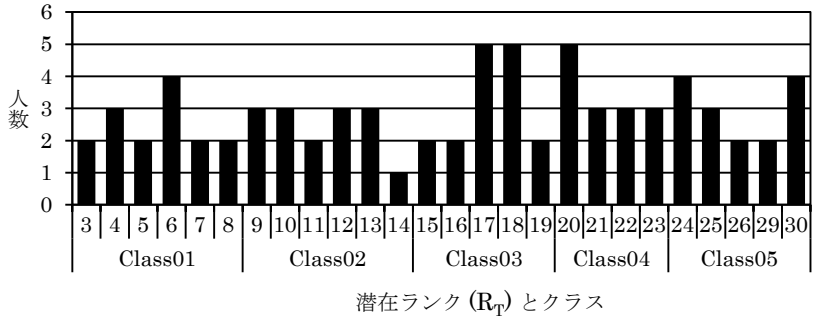

図 35　クラス分けの状況

表 17　各クラスの英語基礎力総合評価 (R_T, θ_T, S_T) の代表値と散布度の比較

Class	*n*	R_T		θ_T		S_T	
		Mdn	*Range*	*M*	*SD*	*M*	*SD*
Class 01	15	6	5	-3.06	0.604	26.9	3.88
Class 02	15	11	5	-1.39	0.584	35.5	3.76
Class 03	16	17	4	0.11	0.652	42.6	3.66
Class 04	14	21	3	0.97	0.698	46.9	3.09
Class 05	15	26	6	2.89	1.204	54.3	4.84

R_Tについてはクラスカル・ウォリス検定, θ_Tと S_Tについては 1 元配置の分散分析で有意差 （p < .01） があり，多重比較を行ったところ，θ_Tについては連続するクラス間以外で，θ_Tと S_Tについてはすべてのクラス間で有意差 （p < .01） が認められた．R と θ の間の相関係数は，3 つの項目の種類ごとに見ても，それらを合計した英語基礎力総合評価で見ても.90 以上の高い値を示した. S と R の相関係数も.90 以上, S と θ の相関係数も.92 以上の高い値を示した（表 18 参照）. これらの状況から，LRT の潜在ランク，1PLM の潜在能力値，正答数，いずれで評価しても，能力はほぼ同じ順位で評価されることがわかる．しかし，正答数による評価は，使用するテスト項目と受験集団に依存した数値であるため，標準化された評価結果として扱うことはできない．一方，LRT の潜在ランクと 1PLM の潜在能力値による評価は，分析を行った事前テストの受験者集団に依存するものの, いったん標準化されたならばそれ以降の集団の特性には依存しない標準化された評価結果である．ただし，1PLM の連続した間隔尺度上で表現される細かい数値は，クラス分けのためには不要であるばかりか，どこに閾値を設けるかの判断が難しい．それに対し，LRT の潜在ランクは，はじめから順序尺度上に限られた数の段階に分けて表現されるので，クラスをどこで分けるかの判断を容易に行うことができる．したがって，プレイスメントテストの分析に LRT を活用することは，妥当かつ有用であると言える（木村，2009a）.

表 18　R, θ, S 間の相関係数 [20]

Vgm	R_{vgm}	θ_{vgm}	S_{vgm}
R_{vgm}	—	.96	.96
θ_{vgm}		—	.99
S_{vgm}			—

Dlg	R_{dlg}	θ_{dlg}	S_{dlg}
R_{dlg}	—	.90	.90
θ_{dlg}		—	.98
S_{dlg}			—

Mlg	R_{mlg}	θ_{mlg}	S_{mlg}
R_{mlg}	—	.93	.92
θ_{mlg}		—	.92
S_{mlg}			—

総合評価	R_T	θ_T	S_T
R_T	—	.96	.94
θ_T		—	.96
S_T			—

7.4.　プレイスメントテストの実施

前節のクラス分けシミュレーションを行った翌年，7.2 節のプレイスメントテストを 125 名の大学入学者に実施し，5 クラスにクラス分けを行った．前節のクラス分けシミュレーションの時には，GNT モデルを処理するプログラムがなかったので，3 つの下位テストの潜在ランク（R_{vgm}, R_{dlg} , R_{mlg}）の単純和（R_T）で総合力を求めクラス分けをしていたが，その後 Exametrika で GNT モデルが扱えるようになったので, 3 つの下位テストの潜在ランク（R_{vgm}, R_{dlg} , R_{mlg}）を求めた後，それらの潜在ランクを項目として，GNT モデルにより，分析するランク数を 5 とし，目標分布を一様分布としてクラス分けを行った．前節の方法との違いを明らかにするため，前節の下位テストの潜在ランクの単純和によるクラス分けも行い，両者の比較を行った．また，前節のシミュレーションのデータに対しても, GNT によるクラス分けを行い, 結果を比較した（木村, 2009b）．

単純和によるクラス分けと GNT によるクラス分けの結果は，いずれの場合も高い順位相関を示した（125 人の実際のクラス分けで 0.95，75 人のシュミレーションデータで 0.93）．しかし，2 つのクラス分けで異なるクラスになったケースも，125 人中 42 人と多かった（シミュレーションデータの場合は 75 人中 4 人）．GNT のクラス分けで単純和のクラスよりも 1 つ上のクラスになった者が 23 人，1 つ下のクラスになった者が 18 人，2 つ下のクラスになった者が 1 人であった（シミュレーションデータの場合は GNT のクラス分けで単純和のクラスよりも 1 つ上と 1 つ下のクラスになった者が 2 人ずつ）．特に 125 人のクラス分けの場合に，2 つの方法で別クラスになることが多かったのは，クラス分けが均等に近くできたかどうかによる影響もあるだろう．それぞれの手法でクラス分けされた人数は表 19 に示すとおりで, GNT によるクラス分けの方が，目標分布を一様分布に指定したこともあり，ほぼ均等に 23～27 人となった．単純和によるクラス分けではばらつきが大きく 20～29 人となった．一方，シミュレーションデータのクラス分け

[20] R_T:θ_T と R_T:S_T はスピアマンの順位相関係数（r_s）を，θ_T:S_T はピアソンの積率相関係数（r）を用いた．

の場合，2 つのクラス分けともほぼ均等（14~16）であったので，2 つのクラス分けで異なるクラスになるケースも少なかった．単純和と GNT の大きな違いは，GNT は下位テストの識別力の差を考慮することである．単純和と GNT によるクラス分けで異なる結果が出るのは、そのことが大きく影響しているだろう。

表 19　異なるクラス分け方法による人数配分の違い

		Class1	Class2	Class3	Class4	Class5
シミュレーションデータ	単純和	15	15	16	14	15
	GNT	15	15	16	14	15
実際のクラス分け	単純和	28	29	20	23	25
	GNT	27	26	26	23	23

単純和の場合，どこにクラスの分け目を持ってきたらよいか，試行錯誤して決めなければならないが，GNT の場合，目標分布を一様分布にしておけば，どこで分けるか悩むことなく，ほぼ均一にクラス分けされる．下位テストの潜在ランクの単純和によるクラス分けよりも，GNT を利用したクラス分けの方が，容易に均等なクラス分けができることが，他の集団に対しても確かめられた（木村，2009c）.

7.5.　多値モデルの分析：Rdg の項目分析

IRT の標準的なモデルにおいて，大問形式の問題に対する応答データを，個々の設問に対する応答を正解・不正解の二値データとして扱うことは，局所独立の仮定を満たさない可能性があるので望ましくない．局所独立の仮定が満たされない場合に，項目パラメータの推定値に影響を与える可能性があるからである．シミュレーション研究（佐野, 2009; 橋本・植野, 2009）でも，局所独立の仮定が満たされない場合に，項目パラメータの推定値が過大評価されたり過小評価されたりすることも報告されている．このことは，間隔尺度上ではなく順序尺度上に能力推定を行おうとする LRT においても同様である．このような場合は，大問ごとの応答を多値の順序データとして扱い, NTT の拡張モデルである段階的ニューラルテスト（graded neural test, GNT）モデル（Shojima, 2007b）によって分析するのが一つの解決策である．本研究では，基礎的な英語読解力を測定することを目的に実施した英語読解問題（1 つの文章に 2～5 の設問）を GNT モデルによって分析した事例を報告する（木村・永岡，2010a）.

問題は実用英語検定協会の許可を得て，英検 3 級・準 2 級・2 級の過去問題を使用し，表 20 に示すように互いに共通項目を含む複数のテストレットを用意し，2009 年 4 月～2010 年 7 月に，大学 1 年生のべ 475 名（うち解答項目数 3 以下の 57 名は分析対象から除外）から解答を得た．集団としては Goup1～Group4 の 4 つで，各集団の読解力に見合っていると授業担当者が判断した問題を基礎力判定テストの一部として利用した．

表 20　テスト項目数と受験者数

グループ	人数	大問記号／小問数／英検級															
		E-1	E-2	E-3	E-4	E-5	F-6	F-7	F-8	G-1	G-2	G-3	G-4	G-5	H-6	H-7	H-8
		2	3	5	3	4	3	4	5	2	3	5	3	4	3	4	5
		3 級			準 2 級		2 級			3 級			準 2 級		2 級		
Group1	132	130	130	127	125	118				121	119	111	104	88			
Group2	73				72	72	72	68	66				71	70	65	58	54
Group3	66				64	64	64	64	63	66	66	66	62	56			
Group4	148												127	146	147	142	126
合計	419	130	130	127	261	254	136	132	129	187	185	177	364	360	212	200	180

各大問の正当数の分布は表 21 に示すとおりであった．3 級と準 2 級の問題は，E-5 を除いてすべて，正当数の一番多いところ(全問正解)に一番多く分布があり，天井効果が見られる．2 級の問題は正当数が中程度のところに分布のピークがきており，全問正解できているのは 5〜15%で多くの者にとって難しい問題であった．

表 21　各大問の正当数ごとの分布

大問記号		E-1	E-2	E-3	E-4	E-5	F-6	F-7	F-8	G-1	G-2	G-3	G-4	G-5	H-6	H-7	H-8
英 検 級		3 級			準 2 級		2 級			3 級			準 2 級		2 級		
正当数	0	4	3	1	6	19	10	17	2	12	3	2	10	13	25	23	22
	1	13	11	4	47	38	58	38	38	36	11	5	51	19	77	49	51
	2	113	31	5	93	58	53	45	41	139	64	17	99	53	81	53	43
	3		85	14	115	71	15	17	28		107	15	204	86	29	45	27
	4			36		68		15	12			52		189		31	28
	5			67					8			86					9

収集した多値データを Exametrika Ver. 4.3（Shojima, K., 2010）を使って，GNT モデルで，SOM のメカニズムで推定を行い，目標潜在ランク分布の指定をせず，単調増加制約もつけずに，潜在ランク数を 2〜10 で分析した．RMP に基づくテストの適合指標（表 22 参照）のうち，AIC ではランク数 6 が， CAIC と BIC ではランク数 3 が最小の値で，もっとも効率のよいモデルであることを示しているが，ランク数 2 と 3 の場合の χ^2 検定の P 値が小さく Fit が棄却されている．総合的に考えて，ランク数 4 が今回のデータ分析に適していると判断した．ランク数 5 でも指標に大きな違いはなく許容されるが，今回の問題が全般的に正答率の高い（難易度の低い）問題が多いことを考えると，今回はランク数 4 に押さえておき，将来的に難易度の高い問題がアイテムバンクに加わったときに，改めて Fit 指標を参照しながら，ランク数を拡張する方が良いであろう．

表 22　RMP に基づくテストの適合指標

ランク数	2	3	4	5	6	7	8	9	10
$\chi 2$	3370.10	2938.62	2758.77	2599.78	2454.17	2348.56	2280.26	2209.46	2161.45
df	2784	2726	2668	2610	2552	2494	2436	2378	2320
P 値	0.000	0.002	0.108	0.553	0.916	0.982	0.988	0.994	0.991
AIC	-2198	-2513	-2577	-2620	-2650	-2639	-2592	-2547	-2479
CAIC	-16217	-16240	-16012	-15763	-15500	-15198	-14858	-14521	-14161
BIC	-13433	-13514	-13344	-13153	-12948	-12704	-12422	-12143	-11841

ランク数 4 で分析した場合の潜在ランクと素点の関係を示すテスト参照プロファイル（test reference profile, TRP）は図 36 のように，各ランクに推定された受験者の分布を示す潜在ランク分布(latent rank distribution, LRD)と母集団の特徴を表現するランク・メンバーシップ分布（rank membership distribution, RMD）を相対度数で示すと図 37 のようになった.

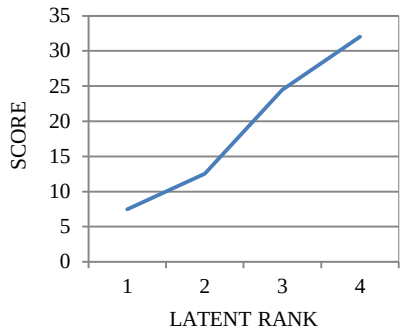

図 36　TRP（ランク数 4）

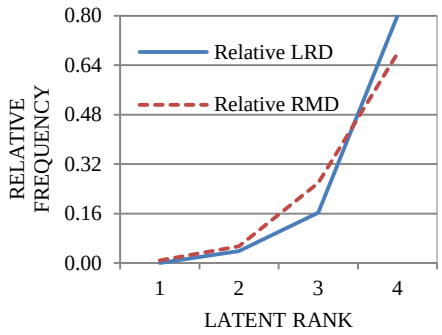

図 37　相対 LRD/RMD（ランク数 4）

相対 LRD から今回の受験者は 80%が R4 であり，R3 と R2 に 16%と 4%，R1 は 0%であることがわかる. 推定されたランクごとの各英検級の平均正解率（表 23 参照）と併せて考えると，今回の受験者集団の 80%は，英検 3 級の読解問題程度の英文を辞書なしでほぼ完全に（91%），準 2 級のものをおおかた（71%）理解できるが，2 級のものは半分程度（53%）しか理解できない. 16%の受験者は，英検 3 級の読解問題程度の英文を辞書なしで 6 割以上理解することができるが，準 2 級のものは少ししか（40%）理解できず，2 級の問題はほとんど理解できない. 4%の受験生は 3 級のものですら十分に理解できない状態であることが分かる. R2 のところで 3 級と準 2 級の平均正答率に逆転現象がみられるのは，R2 のうち 3 級問題を解いているのが 16 人中 3 人だけであるためである（この 3 人は準 2 級の平均正答率も 18%）.

表 23　推定されたランクごとの受験者の各英検級の正解率

潜　在 ランク	度数	相対 度数	平均正答率		
			3 級	準 2 級	2 級
R4	334	0.799	91%	71%	53%
R3	68	0.163	63%	40%	30%
R2	16	0.038	18%	39%	16%
R1	0	0.000	--	--	--

項目の特性を図 38 の IRP で見てみると，英検 3 級と 2 級の問題は 2 種類の問題ともほぼ同じふるまいをしているが, 準 2 級の問題は片方の方はどのランクに推定される受験者もほぼ同程度に正解している様子がわかる．また，それぞれの潜在ランクの受験者が大問中何問正解しているかを示す項目カテゴリ参照プロファイル（item category reference profile, ICRP）を見ると，G-1 のように潜在ランクが上位になるにしたがって，2 問中 2 問正解している割合が増え（2 問中 0 問正解している割合が減る）項目もあるが，G-4 のように潜在ランクが上位になっても，あまり大きく正解率が変化しないものもある（図 39 参照）．前者は識別力の高い項目，後者は識別力の低い項目であるといえる．

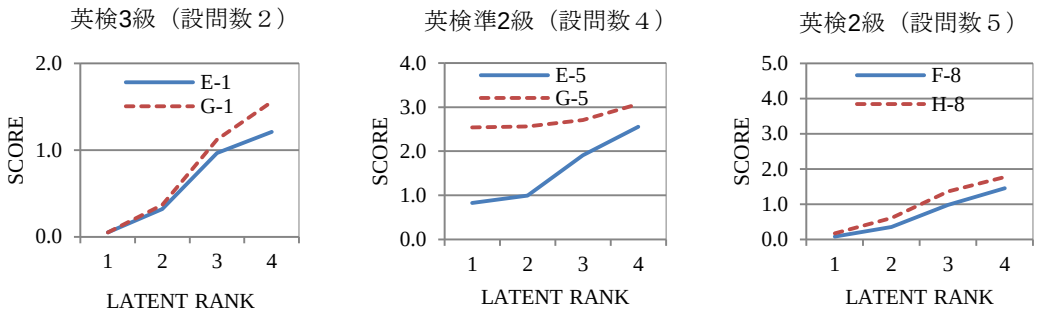

図 38　英検級・設問数ごとの IRP（一部）

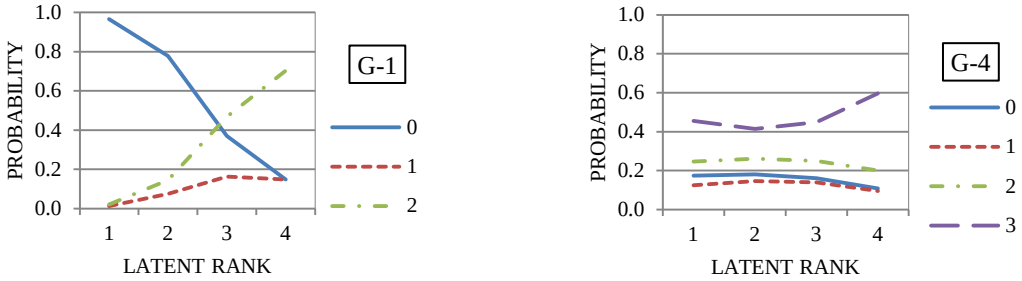

図 39　項目(大問)ごとの ICRP（一部）

大問形式の英語読解問題のデータを GNT モデルにより分析し，潜在ランク上に受験者の能力を推定し，各潜在ランクの受験者の特徴を描き出すとともに，大問形式の問題の特性を分析できることをが示せた．

7.6. アイテムバンクの拡充と項目困難度の等化

第1次事前テスト結果から，項目を固定させたプレイスメントテストに選ばれた項目をアンカー項目として，後に等化してアイテムバンクの項目数を増やすことを目的に，Vgm, Dlg，Mlgの3種類ごとに，4つのテストレットが作成された．

Vgmの場合を例にとって説明すると，この4つのテストは，2009年度に475人の受験協力者に対して実施された．各受験協力者は，2ないし3つのテストを，3ヶ月以上間をあけて受験した（一部に1つしか受験していない受験協力者もいた）．等化の際，複数のテストを受験した協力者のデータは，テスト実施時期が3ヶ月以上あいていたので，別受験者として扱った（のべ1070人）．第２次事前テストの結果から，図40に示すような共通項目による等化デザインで等化を行い，アイテムバンクの項目数は93に増えた（木村・永岡，2010b）．

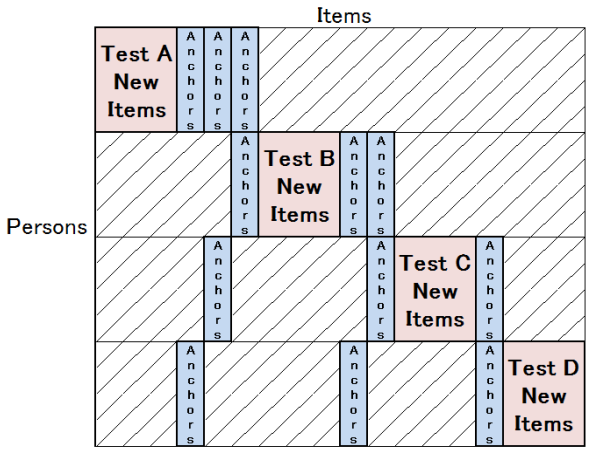

図 40　等化のためのアンカーデザイン

第 3 次事前テストでは，第 2 次事前テストの結果，各アイテムバンクの項目数が増えたので，アンカー項目の見直しを行い，新たに後に等化してアイテムバンクの項目数を増やすことを目的に，6 つのテストが作成された（等化のデザインは第 2 次事前テストと同じ）．それぞれ約 200～300 人の受験協力者からの解答を得て分析を行った（木村，2010a）．

Vgm の場合を例に取ると，アイテムバンクの項目数を 263 に増やした．第２次事前テストのときと同様，複数のテストを受験した協力者のデータは，テスト実施時期が 3 ヶ月以上あいていたので，別受験者として扱った（のべ 1575 人）．第 3 次事前テストの分析が終了した段階での Vmg のアイテムバンクに蓄積された 263 の項目を IRP 指標 β と，項目の出典である実用英語検定の級で整理すると，表 24 のようになる．

表 24　アイテムバンク Vgm の IRP 指標 β と英検級

IRP 指標 β	英検級				合計
	3 級	準 2 級	2 級	準 1 級	
1	37	22	3	4	66
2	8	14	6	3	31
3	3	17	20	8	48
4	1	8	13	8	30
5	0	7	30	51	88
合計	49	68	72	74	263

上記の LRT に基づく分析と並行して，RM に基づく分析も第 2 次・第 3 次事前テストの結果に関して，順次アイテムバンクごとに行われた．ただし，次節で述べるミスフィットの基準の見直しが途中で行われ，第 1 次事前テストの分析に戻って再分析が行われた．その結果について，2 値モデルの Vgm，Dlg，Mlg，について示すと表 25～表 27 のようになる．

表 25　アイテムバンク Vgm の RM に基づく項目困難度基本統計量

	英検級				合計
	3 級	準 2 級	2 級	準 1 級	
N	49	67	69	73	258
M	-1.41	-0.47	0.52	1.57	0.19
SD	0.80	0.91	0.81	0.84	1.37
Max	0.43	2.00	3.19	3.50	3.50
Min	-3.38	-3.11	-0.64	-0.29	-3.38

表 26　アイテムバンク Dlg の RM に基づく項目困難度基本統計量

	英検級				合計
	3 級	準 2 級	2 級	準 1 級	
N	40	38	57	44	179
M	-1.25	-0.01	0.82	1.26	0.29
SD	1.47	1.02	1.13	1.42	1.57
Max	2.35	2.51	3.57	4.05	4.05
Min	-4.45	-2.44	-1.11	-2.25	-4.45

表 27　アイテムバンク Mlg の RM に基づく項目困難度基本統計量

	英検級				合計
	3 級	準 2 級	2 級	準 1 級	
N	40	37	52	---	129
M	-0.56	0.71	0.72	---	0.32
SD	1.08	0.95	1.08	---	1.20
Max	2.4	2.87	3.41	---	3.41
Min	-2.73	-0.71	-1.54	---	-2.73

7.7.　RM におけるミスフィットの基準見直と再分析

第1次事前テストと第2次事前テストでは，CTTのIT相関が0.25未満の場合と，RMで示されるフィット指標*Zstd*の値が1.96以上のものをすべて除去していたが，*Zstd*の値はサンプルサイズが大きくなると過剰に反応する傾向があるので，ミスフィットの判断方法を見直し，これに$Outfit\ MNSQ$の値を加えることにした（Kimura & Nagaoka, 2010）．$Outfit\ MNSQ$の判断基準についてはいくつか提案されているが, 本研究ではBond & Fox(2007)で紹介されている基準1.3とした．つまり，旧判断基準は,

$$\textbf{IT 相関}<\textbf{0.25} \quad \text{かつ} \quad \boldsymbol{Zstd}>\textbf{1.96} \tag{39}$$

であったものを，次のように変更することにした．

$$\boldsymbol{Outfit\ MNSQ}>\textbf{1.3} \quad \text{かつ} \quad \boldsymbol{Zstd}>\textbf{1.96} \tag{40}$$

結果として，サンプルサイズが大きいために*Zstd*値が高くなってしまう場合でも，$Outfit\ MNSQ$の値も参照するので，$Outfit\ MNSQ$の値が小さい項目が除去されなくなった．

また，事前テストを分析する際に，受験者がでたらめに解答して，その受験者にとって非常に難しい問題に偶然正解してしまうケースや, その受験者にとっては非常に易しい問題に不正解してしまうケースによって，項目のミスフィット指標が高くなり排除されてしまう可能性がある．項目のミスフィットと同じ方法で受験者のミスフィットを判定してその受験者を分析対象から除外する方法や他の指標もいくつか提案されているが（Drasgow et al, 1987; Meijer & Sijtsma, 1995; Egberink et al, 2010），テストの一部の反応の異常さによって，その受験者のすべての反応を除去するのは効率が悪い．これを回避するために，上記基準で行われた分析結果に基づき，各受験者の能力（θ_i）と各項目の困難度（b_j）をもとに，次の２つのケースについて，その解答だけ無解答と扱い，再分析を行うこととした．

$$\boldsymbol{\theta_i - b_j < -2} \quad \text{または} \quad \boldsymbol{\theta_i - b_j > 3} \tag{41}$$

理論編3.3で示した表3から分かるように，(41)式の前半の条件は正答確率が12%以下である状況であり，(41)式の後半の条件は正答確率が95%以上の状況であるので，それぞれ，偶然の正解と，偶然の不正解（うっかりミス）と考えられる．式（39）で示したこれまでのミスフィット基準を基準（1），式（40）で示した新しいミスフィット基準を基準（2），基準（2）に式（41）で示される基準を加えたものを基準（3）とする．以前，第1次事前テストのVgmの80項目の分析に関して，基準（1）で行った結果と，基準（2）で再分析した場合と，基準（3）で再分析した場合を比較した．基準（2）あるいは基準（3）を導入することで，事前テストの段階でミスフィット項目として除外される項目は少なくなり，各アイテムバンクに残った項目数は表28に示すように多くなった．

表 28　ミスフィット基準の見直しで各アイテムバンクに残った項目数

	実施した 項目数	基準（1） で残った項目数	基準（2） で残った項目数	基準（3） で残った項目数
Vgm	80	32	71	79
Dlg	47	13	46	47
Mlg	35	19	35	35

3通りの基準でミスフィットを除去して分析された項目困難度がどの程度一致しているか，Vgmについて，すべての分析で残った32項目の項目困難度の相関係数を調べると，0.996から0.998の非常に高い相関がみられた．ただし，表29に示すように，項目困難度の基本統計量を調べると，平均値と最大値・最小値は，ミスフィット除去の基準の違いによって，少し異なっている．これは，各分析でアイテムバンクに残った項目数が大きく異なるためであろう．

表 29　項目困難度の基本統計量

	基準（1）	基準（2）	基準（3）
M	-0.48	-0.23	-0.36
SD	0.94	0.92	0.91
Max	1.61	1.91	1.75
Min	-2.15	-1.80	-1.82

最初に定めた基準(1)を改め，基準（3）でミスフィットを判断することとし，すべてのアイテムバンクについて再分析を行い，第2次・第3次事前テストの分析と等化も基準(3)で行うこととした．

7.8. RM におけるアイテムバンクの統合

論理的に考えて，ダイアローグを聞いて解答する聴解問題 Dlg と，モノローグを聞いて解答する聴解問題 Mlg が，異なる能力を測っているとは考えにくく，各テストの結果においても，Dlg と Mlg の得点の相関は高い．別々のアイテムバンクとして扱うよりも，１つのアイテムバンクに統合して扱う方が，合理的であり，将来的に CAT を運用する場合も扱いやすいので，１つのアイテムバンクに統合することを試みた（木村・永岡，2011b）．

本研究では，2008 年度から 2009 年度に大学 1 年生に対して実施したテストにおいて，Dlg と Mlg の両方のテストを受験している受験者 1046 人を対象に，共通受験者による等化分析 (separate-estimation common-person test equating)を行った．

この受験者が解答している項目は，Dlg が 51 項目，Mlg が 45 項目であった．それぞれの基本統計量は表 30 に示すとおりであった．Measure は項目困難度，S.E.は測定誤差を示す．

表 30　Mlg と Dlg の項目困難度基本統計量

	Mlg		Dlg	
	Measure	*S.E.*	*Measure*	*S.E.*
N	51	--	45	--
M	0.48	0.67	0.36	0.46
SD	1.26	0.19	1.07	0.07
Max	5.14	1.98	4.28	1.09
Min	-4.11	0.56	-3.59	0.40

平均も標準偏差も異なるので，Mlg の尺度上に Dlg の *Measure*（項目困難度）を合わせるために次の式により Dlg の *Measure*（項目困難度）を変換することで等化した．

$$（\mathrm{Dlg} - 0.36）\div 1.07 \times 1.26 + 0.48 \qquad (42)$$

等化した結果を図示すると図 41 のようになった．ほぼすべての受験者を示すプロットが，傾向線の上下に描かれた 95%の信頼区間を示す２つの近似曲線の内側にあることから，Mlg と Dlg は同じ能力を測定しているものと考え，1 つのアイテムバンクに統合して問題ないと判断される．また，Mlg と Dlg の *Measure* の相関係数 0.89 から判断しても同様の結論に至る．

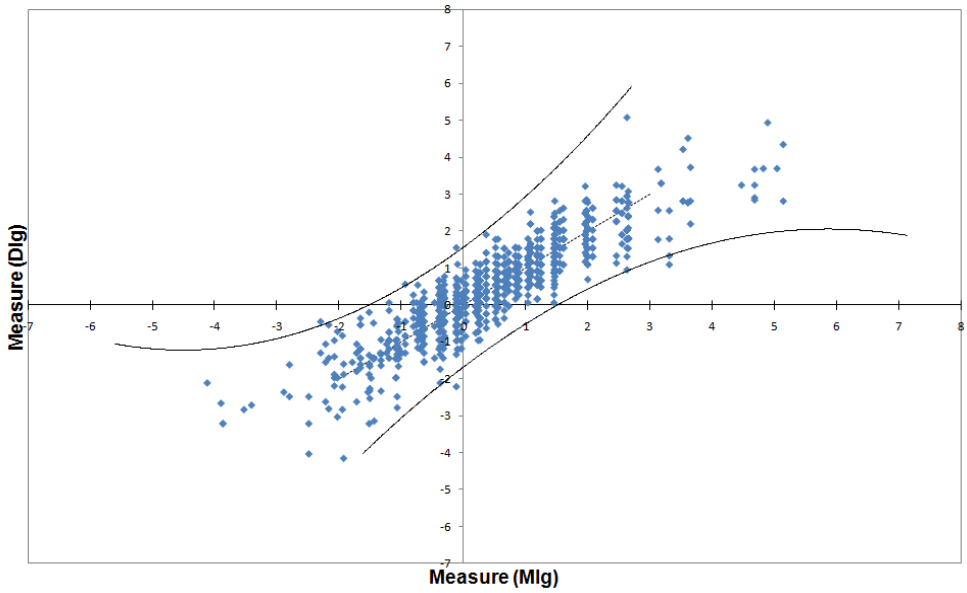

図 41　Mlg *Measure* と Dlg *Measure* の統合

上記分析対象とならなかった Dlg の項目についても，上記変換式で Mlg の尺度に合わせ統合し，聴解力を表す 1 つの尺度としてとらえ，今後は Dlg と Mlg のアイテムバンクは 1 つの聴解力のアイテムバンク Lng として扱うことにした．また，LRT による分析についても，同様に Dlg と Mlg は 1 つの聴解力のアイテムバンク Lng として統合する．統合されたアイテムバンク Lng の RM に基づく項目困難度の基本統計量を，英検の級ごとに細分化したものとあわせて示すと，表 31 のようになる．

表 31　アイテムバンク Lng の RM に基づく項目困難度基本統計量

	英検級				合計
	3 級	準 2 級	2 級	準 1 級	
N	80	75	109	44	308
M	-0.90	0.35	0.77	1.26	0.30
SD	1.33	1.05	1.11	1.42	1.43
Max	2.40	2.87	3.57	4.05	4.05
Min	-4.45	-2.44	-1.54	-2.25	-4.45

8.　CAT 開発フレームワーク第 4 段階での実践的研究

CAT 開発フレームワーク第 4 段階での実践的研究については，シミュレーションにより CAT の仕様の検討と，実テストによるアイテムアバンクの検証について報告する．

8.1. シミュレーションによる LRT-CAT 仕様の検討

次節で述べる LRT-CAT を実施する前に，シミュレーションにより 2.5 節で提案したアルゴリズムでの真値がどの程度再現されるかを検証するとともに，実際の CAT を実施時に指定する終了項目数を検討した．

8.1.1. シミュレーションに使用した IRP と RMP

シミュレーションに使用したIRPは，7.6節で述べたアイテムバンクの263項目のIRPである．既知の受験者情報として使用したRMPは，このアイテムバンクを拡充させるために，事前テストでIRPを調べてあるアンカー項目6項目を含む26項目からなる6種類の第3次事前テスト（2010A～2010F）の受験に協力した延べ1575人のテスト結果のRMPで，これを真値とした．いずれも，Exametrika 4.3 （Shojima, 2010）を使い，SOMによる推定法で，潜在ランク数5，事後分布の指定なし，IRPの単調増加制約なしで分析したものである．

IRP指標βの分布と受験者の潜在ランクの真値の分布は，図42と図43に示すとおりである．図42では, 項目が英検のどの級の項目であるかも併せて示したもので表24をグラフ化したものである．

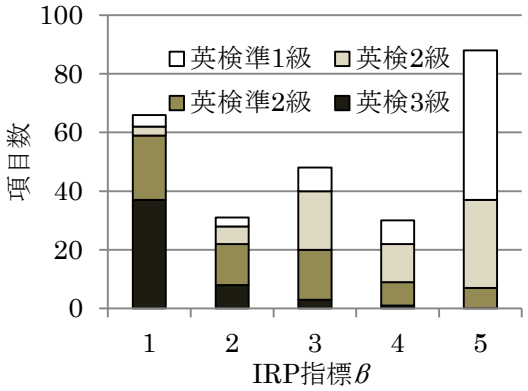

図 42　IRP 指標の分布（n=263）

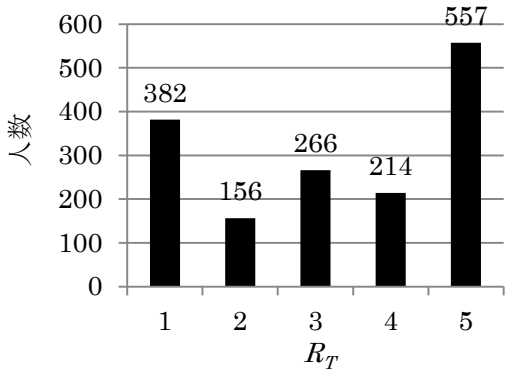

図 43　R_Tの分布（N=1575）

8.1.2. シミュレーション条件

各潜在ランクから100人をランダムに抽出した500人のRMPを使い，10回ずつ2.5節で提案したLRT-CATアルゴリズムによりシミュレーションを実行した．正答･誤答の判断は，0から1の間の一様乱数を発生させ，(43)式（RMPとIRPの積和）によって求められる受験者iの項目jに対する正答確率（p_{ij}）と比較し，等しいかp_{ij}の方が大きい場合に正答，小さい場合に誤答とした．

$$p_{ij} = \sum_{q=1}^{Q} p_{iq} \upsilon_{jq} \tag{43}$$

そのほかのシミュレーション条件は，これまでの議論を踏まえた上で，以下のように設定した．

(1) 潜在ランク数（Q）：5

(2) 初期能力値：初期RMPが一様分布

$$\mathbf{p}_i^{(0)} = [0.2, 0.2, 0.2, 0.2. 0.2]$$

(3) 推定方法：最尤推定

(4) 項目選択：最初はすべてのIRP指標βから1項目ずつランダムに抽出し，5項目からなるテストレットとして実施し，暫定RMPを推定．それ以降は，IRP指標βが暫定の潜在ランクの推定値±１の範囲の未使用項目から，λの値が最小となる項目を1項目選択実施し，暫定RMPの推定を繰り返す．

(5) 終了条件：実施項目数mが「m>20」かつ，RMPの変化量を示すμが「μ<0.05」

8.1.3. シミュレーション結果

(1) 潜在ランクの真値の再現性

シミュレーションによって得られた5000件の潜在ランクの推定値（$\hat{R}$）の分布を潜在ランクの真値（R_T）ごとに示すと，表32のようになった．各潜在ランクから同数抽出しているので，理論的に$\hat{R}$は一様分布に近くなるはずだが，$\hat{R}$=2で少なく（一様分布から期待される人数（1000人）より－24%），$\hat{R}$=5で多くなっている（一様分布から期待される人数（1000人）より＋22%）．

R_Tごとに見ると，R_T=1とR_T=5の1000件は，シミュレーションで74%と73%の割合で真値と同じランクに推定されているが，R_T=2，R_T=3，R_T=4の1000件は，シミュレーションで46%，58%，44%しか同じランクに推定されていない．潜在ランクの真値の再現性はあまり高いとは言えない．しかし，本研究のCATアルゴリズムでは潜在ランクではなく，RMPを使っている．潜在ランクの真値の再現性の低さの原因を考えるには，今回のシミュレーションで真値として使用した事前テストの受験協力者の推定RMPの状態を分析してみる必要があるだろう．

表 32　R_T ごとの $\hat{R}$ の度数分布

		R_T					合計
		1	2	3	4	5	
$\hat{R}$	1	744	257	3	0	0	1004
	2	210	459	94	1	0	764
	3	32	230	578	172	21	1033
	4	12	44	229	444	246	975
	5	2	10	96	383	733	1224
合計		1000	1000	1000	1000	1000	5000

(2) $\hat{R}$ と R_T の一致度

シミュレーションにより得られた $\hat{R}$ と R_T の差がどの程度のものが, どのぐらいの頻度であるかを調べたところ表33のようになった．両者が一致したものは5000件中59.2%，$\hat{R}$ が R_T より１ランク上に推定されたケースが21.0%，1ランク下に推定されたケースが15.4%であった．$\hat{R}$ -R_T が4となったケースが2件，3となったケースが22件あったが，－4や－3となったケースはなかった．

表 33　$\hat{R}$ と R_T の一致の程度

$\hat{R}$ -R_T	頻度	%	
4	2	0.0%	3.9%
3	22	0.4%	
2	172	3.4%	
1	1052	21.0%	95.6%
0	2958	59.2%	
-1	769	15.4%	
-2	25	0.5%	0.5%
-3	0	0.0%	
-4	0	0.0%	

全体の95%以上が±1の範囲に収まっているものの，1ランクのずれが生じたものが上下合わせて36.4%あり，安定しているとはいえない．その原因を探るためにも，今回シミュレーションに使用した事前テストの受験協力者の推定RMPを分析してみる必要がある．

(3) シミュレーションに使用した事前テストの受験協力者の推定 RMP

今回のシミュレーションに選ばれた500件のRMPの特徴を概観するために，R_T ごとにRMPを平均してグラフ化したのが図44である．図44から分かるように，　R_T が潜在ランク両端のRMPは，

R_Tが両端以外の場合に比べて，平均的にピークがはっきりしており，かつピークの値が高い．換言すれば，今回使用したRMPの平均は，$\hat{R}$が両端の場合でRMPベクトルの中で最も大きな値は比較的高い（0.63と0.51）が，両端以外の場合はRMPベクトルの中で最も大きな値は0.40前後であり，両端以外の場合，RMPベクトルの中で2番目と3番目に大きい値を合計すると0.42～0.57もある．

すなわち，今回のシミュレーションで潜在ランクの真値の再現性が，潜在ランクの両端で高く，両端以外で低かったのは，真値として使用したRMPが両端ではピークが明確で，R_T=1とR_T=5のピークは0.63と0.51と高いのに対して，両端以外のピークはR_T=2，R_T=3，R_T=4で，それぞれ0.42，0.42，0.37と低く，R_T=2とR_T=4では，それぞれ隣接する潜在ランクR_T=1とR_T=5との差が小さいことに原因があると思われる．$\hat{R}$とR_Tの一致度を検討した際に見られた偏りも，同じ原因によるものだと考えられる．

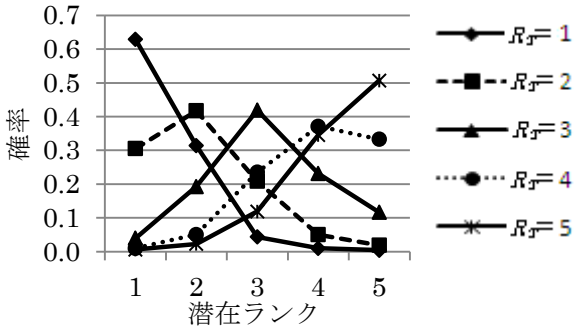

図 44　R_T ごとの RMP の平均

(4) ランク・メンバーシップ分布

LRTにおいて，母集団の周辺分布の特徴を表すものとして，ランク・メンバーショップ分布（rank membership distribution, RMD）という概念がある．RMDは，各受験者が各潜在ランクに所属する確率を示すRMPの単純和であり，受験者の母集団の周辺分布の特徴を表す．

和を1.00に調整された相対RMDを調べると，R_Tも$\hat{R}$も，いずれの潜在ランクについても，その値は0.190から0.209の間に収まり，ほぼ一様分布であった（表34参照）．このことは，今回のシミュレーションにおいて，前項で見たように潜在ランクの再現性という意味では，不安定さを見せたが，母集団の周辺分布に関しては，ほぼ再現されていたことを示すものと考えられる．

表 34　相対 RMD の再現性

	1	2	3	4	5
R_T	0.206	0.190	0.209	0.197	0.199
$\hat{R}$	0.199	0.201	0.203	0.199	0.198

(5) 終了項目数

上記終了条件を満たすのに要した項目数については，予想していたよりもかなり多かった．特に $\hat{R}$ -R_Tが±1の範囲の場合，平均で約25項目と他に比べて多く，標準偏差も他に比べて大きい．最大で55~58項目を実施するまで，終了条件のμ<0.05を満たすことはできなかった（表35参照）．

$\hat{R}$ ごとの終了項目数を見ると，$\hat{R}$ =2が最も終了項目数が多くなっている（図45参照）．これは，本研究で使用しているアイテムバンクにIRP指標β=2である項目が少ないことが影響しているのではないだろうか（図42参照）．このことは，現アイテムバンクはIRP指標β=2である項目を増やす必要があることを示唆している．言い換えると，現アイテムバンクでLRT-CATを実施しても，真の潜在ランクが2の受験者の能力を効率よく測定することが困難であるといえる．

終了項目（m）ごとにRMPの真値がどのような特徴を持っていたか調べると，図46に示すように，mが多くなるにつれて，RMPのR_I=2の割合が大きくなることがわかる．mが28以下の場合に比べて，mが29以上の場合，R_I=2の割合が高い．このことからも，現アイテムバンクでLRT-CATを実施しても，真の潜在ランクが2の受験者の能力を効率よく測定することが困難であり，そこが現アイテムバンクの弱点であることがわかった．

表 35　$\hat{R}$ -R_T ごとの終了項目数

$\hat{R}$ -R_T	*M*	*SD*	*Max*	*Min*
4	22.0	1.00	23	21
3	21.5	0.94	25	21
2	22.3	2.84	37	21
1	24.9	5.29	58	21
0	24.8	5.30	56	21
-1	25.0	4.99	55	21
-2	24.5	3.92	38	21

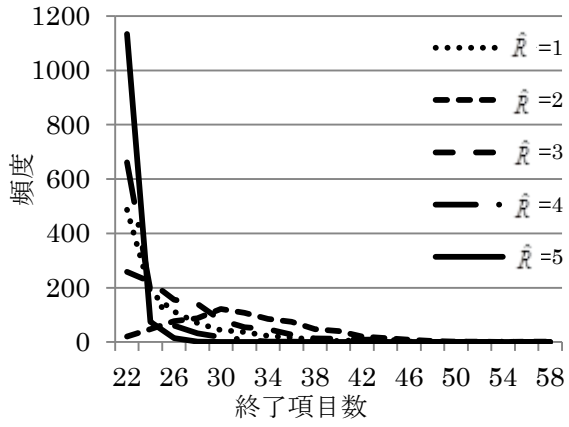

図 45　$\hat{R}$ ごとの終了項目数

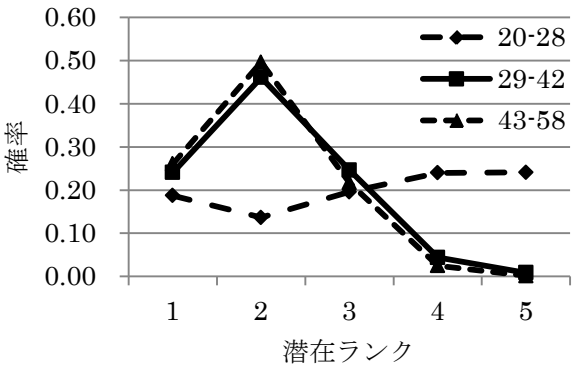

図 46　終了項目数と RMP 真値平均

しかし, CATを実装するのに常に理想的なアイテムバンクを維持することは極めて困難である. 現状でのアイテムバンクの弱点とCATによる能力推定の限界を把握した上で, 現段階のアイテムバンクでCATを実装するとしたら, 何項目実施すればよいかを探ることが現実的であると考える. 今回のシミュレーションの結果から，IRP指標β=2である項目が現アイテムバンクの弱点であり，$\hat{R}$=2の受験者の能力推定に限界があることがわかるので，それ以外のランクの受験者について，何項目実施すればよいかを検討することにした.

シミュレーションデータから$\hat{R}$=2のものを除いて集計しなおすと，終了項目数28までで，約90%が終了条件を満たしていることがわかる（図47参照）.

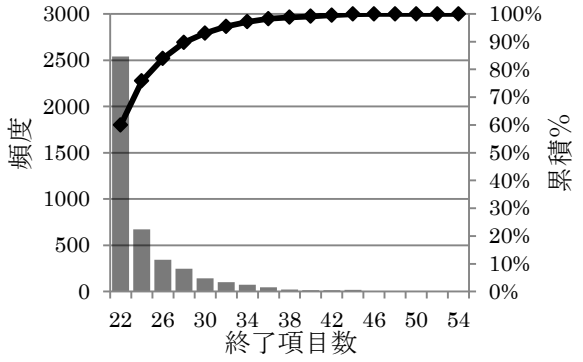

図 47　$\hat{R}$=2 以外の終了項目数

8.2.　LRT-CAT を使った実テストによるアイテムバンクの検証

前節のシミュレーションの結果をもとに，実際に LRT-CAT でテストを実施する条件（指定項目数）を決定し，実テストにより Vgm のアイテムバンクの検証を行った.

8.2.1. LRT-CAT の実施条件

前節で述べたシミュレーションの結果から, 現在のアイテムバンクでLRT-CATを実施した場合, 潜在ランクが2の能力の受験者を測定するには，IRP指標β=2である項目が不足しているため，他の潜在ランクの受験者より精度が低いことが分かった．しかし，それ以外の潜在ランクの受験生に対しては，28項目で約90%の受験者のRMPの変動がほぼ収束する（μ<0.05）ことがわかったので，終了条件を「m=28」として実施することとした．

8.2.2. LRT-CAT の受験協力者と実施手順

大学1年生180人が英語授業の課題の一部として, 今回のLRT-CATを受験した．テストはMoodle（Ver. 1.9）上で行われた．第3章で述べたLRT-CATアルゴリズムをMoodle上で実施するエンジンは，Moodle上でLRT-CATを実装するモジュール（秋山・木村・荘島, 2011）で行った．前節で述べたシミュレーションも，同モジュールの一部の機能を使って行われた．

学生は決められた期間内の好きな時間に，指定のMoodleにアクセスし，テストを1回だけ受験した（2回以上受けられない設定にした）．学生に対して，このテストはコンピュータが出題する問題の難易度を調整するCATであり，受験する者によって異なる問題が出題されること，異なる問題を解いても結果は統計的に調整され公平に評価されること, 特に時間制限は設けないが必ず最後の問題まで解答することなどが事前に教場で説明された．

解答開始時刻と終了時刻を記録し, どのくらい時間をかけてテストに取り組んだか分かるようにした．各受験者にアイテムバンクのどの問題がどのような順番で出題されたか, それぞれの問題に正解したかどうかも記録された．また，全問解答後には，画面にRMPを提示した．RMPをどのように解釈したらよいかについても，事前に授業内で説明を行った．

8.2.3. LRT-CAT の結果

LRT-CATで28問すべてに解答した180人のうち，所要時間が短すぎる5分未満の4人と，長すぎる50分以上の16人のデータを分析から除外した．160人分の受験結果を$\hat{R}$ ごとに見ると，$\hat{R}$ =5が49人（31%）で一番多く，$\hat{R}$ =3，$\hat{R}$ =1，$\hat{R}$ =4，$\hat{R}$ =2の順で，$\hat{R}$ =2が16人（10%）で一番少ない（図48参照）．この $\hat{R}$ の分布は，図43に示したシミュレーションで真値として利用した第3次事前テストの受験協力者1575人のR_Tの分布によく似ている．両集団は異なる年度だが, いずれも大学1年生であり，複数の同じ大学の同じ専攻の学生であることから，十分納得できる．

RMPの推定がどの程度収束していたかについて，最終μの値をみると，約74%が0.05未満に収まっているが, 残りの25%は0.05以上であり，最大は0.14で，0.10以上のケースは約4%あった（図49参照）．

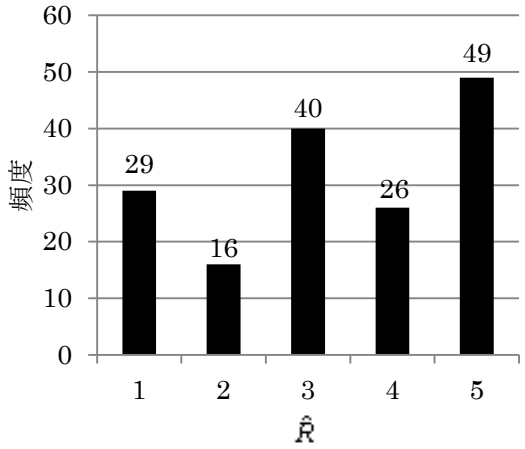

図 48　受験結果：$\hat{R}$ 別人数

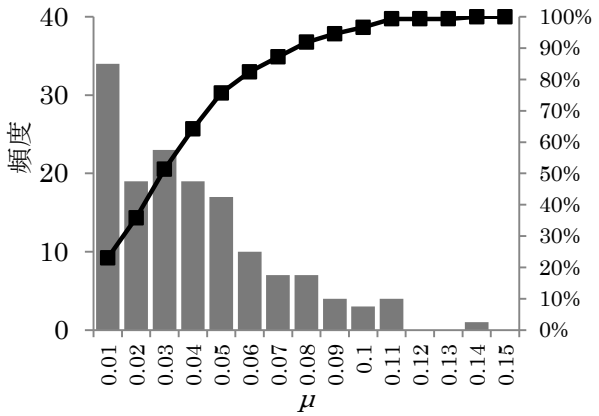

図 49　μ の分布

$\hat{R}$ 別に整理しなおすと，0.10 以上のケースは $\hat{R}$ =2（3 件）と $\hat{R}$ =3（3 件）の場合にだけ見られた．これは，先に述べたように，現在のアイテムバンクに $\hat{R}$ =2 の能力の受験者を測定する項目（IRP 指標 β=2 である項目）が不足しているためだと考えられる（図 50 参照）．この潜在ランクが 2 の受験者を測定する項目（IRP 指標 β=2 である項目）が不足している影響は，隣接する潜在ランクが 3 の受験者にも表れているように思われる．μ が 0.05 未満にならないケースは，$\hat{R}$ =5 は 1 件（2%），$\hat{R}$ =1 と $\hat{R}$ =4 は 6 件（21%と 23%）であるのに対して，$\hat{R}$ =2 と $\hat{R}$ =3 でそれぞれ 13 件（81%）と 16 件（40%）と多くなっている．

実テストの結果からも，前節のシミュレーションによる分析結果と同様，現在のアイテムバンクの状況では，R2 レベルの能力を効率よく判定できないことが分かった．その主な原因はIRP指標βが2の項目がアイテムバンクに少ないことであり，このレベルの困難度の項目をアイテムバンクに追加すべきであることが示唆される．

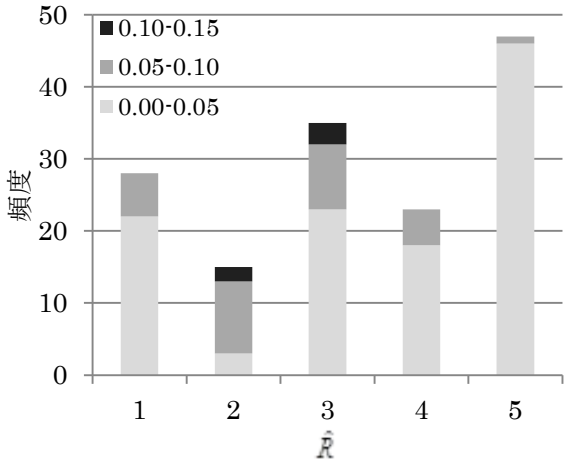

図 50　潜在ランク別最終 μ

8.3.　　Moodle UCAT を使った実テストによるアイテムバンクの検証

Vgm のアイテムバンクと Lng のアイテムバンクについて，RM による実テストを実施し，両アイテムバンクの検証を行った（木村・永岡，2012c）．Vgm アイテムバンクの項目の状況は 7.6 節の表 25 に，Lng のアイテムバンクの状況は 7.8 節の表 31 に示したとおりである．これを困難度ごとにどのような項目がそれぞれのアイテムバンクに蓄積されているかを図示したものが, 図 51 と図 52 である．

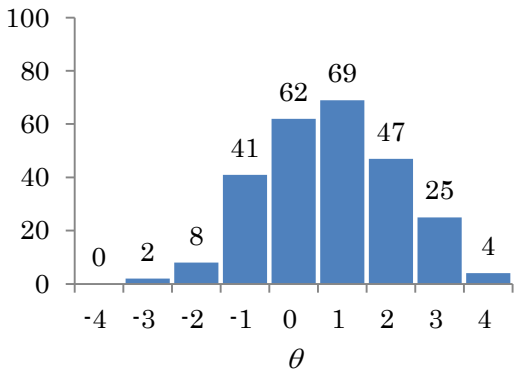

図 51　困難度ごとの項目数（Vgm）

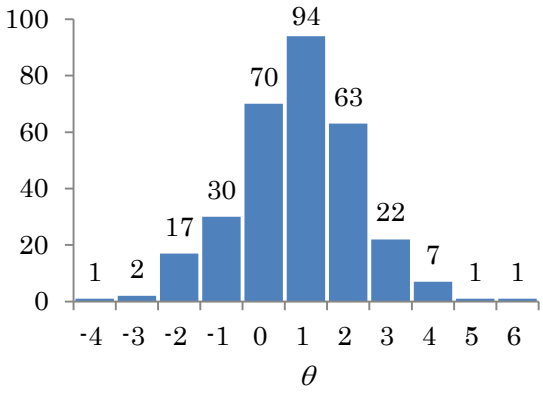

図 52　困難度ごとの項目数（Lng）

CAT の終了条件は，理論的に S.E.が 0.5logits になる項目数 16 をとして（Linacre, 2006），約 160 人の大学 1 年生に，Moodle UCAT（Kimura, Ohnishi, Nagaoka,2012）を使って，Vgm と Lng 別々に実施した. 分析は, 途中で受験を放棄した者, 所要時間が短すぎる者・長すぎる者を除外して,

145 人分の Vgm の CAT データと，130 人分の Lng の CAT のデータを対象に行った．いずれの CAT も 90%以上が S.E. 0.55logits 以下で終了していた．

CAT が終了時た時点で各アイテムバンクの項目の使用頻度を調べた．今回の CAT でどのレベルの項目がどの程度の頻度で使用されたか, 受験者 100 人当たりの項目困難度別の平均使用頻度をまとめたものが，図 53 と図 54 である．

図 51 から分かるように，Vgm のアイテムバンクの方は，困難度が-2～-1logit 程度の問題が現アイテムバンクで少なく, 図 53 に示されているように, 今回の CAT の使用頻度から困難度が-3.0～-1.0 程度の項目の平均使用頻度が他の困難度よりも高く，このあたりの困難度の項目が不足していることがわかる．また，図 52 から分かるように，Lng のアイテムバンクの方は困難度が-4.0～-3.0 程度の問題が現アイテムバンクに 3 項目しかなく，図 54 に示されているように，今回の CAT の使用頻度から見ると，-3.0 程度の項目の平均使用頻度が 7.6 回と非常に高く，このあたりの困難度の項目が非常に不足していることがわかる．

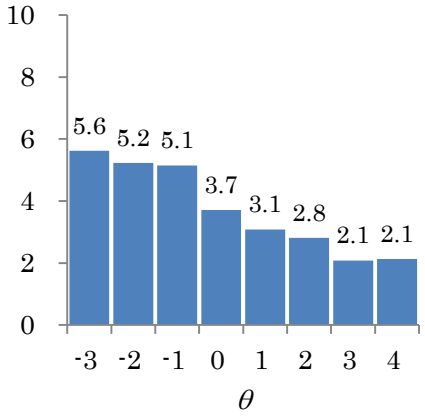

図 53　100 人当たりの項目使用頻度（Vgm）

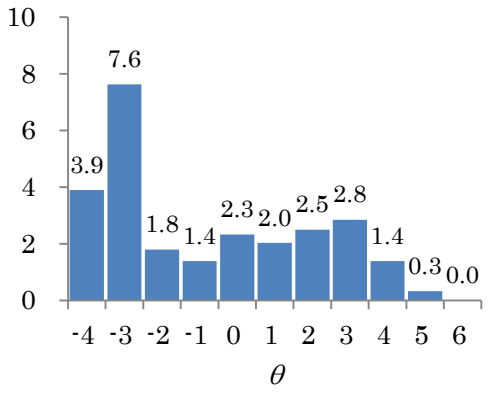

図 54　100 人当たりの項目使用頻度（Lng）

Vgm のアイテムバンクについては，困難度が-3.0～-1.0 程度の項目が不足気味であり，8.2 節の LRT による分析結果と同様，困難度の低めのところに弱点があることが示唆されている．

どのような目的でこのアイテムバンクを使って CAT を実施するかにもよるが，今回の CAT 受験者と同等の受験者に受験させ，幅広く能力の差異を見極めるのであれば，いずれの理論により CAT を実装するとしても，困難度の低めの問題を追加する必要がある．しかし，もし CAT の目的が，一定レベル（たとえば LRT における潜在ランク 3，あるいは RM における 1logit）に達していない受験者を見極めることが目的であるなら, 現状のアイテムバンクで十分であるともいえる．

Lng のアイテムバンクについては，RM による分析結果から，困難度が 1logit 前後(0～2±1logit)の項目に 73%が集中しており，今回の受験者層に対しては，それよりも易しい項目が不足していることが，CAT を実施して得た困難度別の項目平均使用度から明らかになった．今回の CAT

受験者と同等の受験者に受験させ，平均よりも下の能力の差異を識別するためには，困難度の低めの項目をアイテムバンクに追加すべきである．しかし，Vgm と同様，一定レベル（たとえばRM における 1logit）に達していない受験者を見極めることが目的であるなら，現状のアイテムバンクで十分であるともいえる．

9. CAT 受験者に対するアンケート調査

CAT を受験した後の学生の反応は，2.3 で述べたように，どのレベルの受験生も 50%ぐらいしか正答できないために，心理的に落ち込んでいることが多い．本章では，CAT 実施後に受験者に対して行った２つのアンケート調査に基づき，CAT の心理学的側面について考察を加える．

9.1. CAT に対する大学生の一般的な反応

Kimura & Nagaoka (2011b) は，一般的な項目選択ルール（情報量を最大化するために目標正答確率を50%にするルール）により実施されたCATの受験者にアンケート調査を行い，受験したCATに対する難しさの印象，CAT受験経験に対する心理的印象（幸福感）を調べるとともに，受験したCAT全体でどのくらい正答できたと感じたか尋ねた．また，これらの回答について考察を加えるために，高校の英語のテストでどのくらい正答できていたか，一般的にテストでどのくらい正答できることを期待するか，どのくらい正答できると満足（あるいは不満足）と感じるのか，についても回答を求めた．併せて，CATの仕組み（項目困難度の調整）に気づいたか，異なる問題に回答することを不公平に感じるかについても尋ねた．

具体的にはCATを受験した大学生の一般的な反応を探るために作成した下の9項目からなるアンケートを，CAT（CASEC : 英語力を測定するCAT）を受験した156人の大学新入生にを対象に，CAT受験の翌週に実施した．なお，ほぼ全員がCAT受験は初めてであるので，CAT実施前にCATの仕組み（出題される順番に問題に解答しなければならないこと，後から前の問題の解答を変更できないこと，受験者の解答によって次の問題の難易度が調整され出題されること，したがって，解答する問題の種類と数は同じだが同じ問題を解くわけではないこと，異なる問題に解答しても全員のテスト結果は比較可能なスコアとして示させること）については説明を行った．

Q1. CASECは難しかったですか（易しかったですか）.
Q2.自分の回答によって次の問題の難易度が調整されていると感じたことはしますか.
Q3. CASECは受験者ごとに出題される問題が異なります. 異なる問題が出題されるのは不公平だと感じますか.
Q4. CASECを受験し終わった時，うれしい（幸せな）気持ちになりましたか，悲しい（不幸な）気持ちになりましたか.
Q5. CASECは100点満点で結果は出ませんが，自分の得点を100点満点で評価すると今回のCASECの結果は何点ぐらいに思えますか.

Q6. 高校の時に学校で受けた定期試験の英語のテストは100点満点でだいたい何点ぐらいでしたか.

Q7. 一般的にあなたがテストを受けるときに期待する得点は100点満点で何点ぐらいですか.

Q8. 一般的にあなたがテストを受けて，満足する点数は100点満点で何点ぐらいですか.

Q9. 一般的にあなたがテストを受けて，がっかりする点数は100点満点で何点ぐらいですか.

Q1「難しさに」については，82人（53%）が「とても難しい」，60人（38%）が「少し難しい」と回答しており，両者を合わせると，90%以上の学生がCASEC(CAT)を難しいと感じていた．Q2「難易度が調整されていることへの気づき」については，「しばしばあった」と「ときどきあった」を合わせて83人（53%）が気づいていたが，「ほとんど気づかかった」または「まったく気づかなかった」という回答も73人（47%）であった．Q3「問題が異なることへの不公平感」については，「強く感じる」と「少し感じる」を合わせても19人（12%）で，「どちらともいえない」が35人（22%），「あまり感じない」と「まったく感じない」をわせると102人（65%）であった．Q4の「受験後の気持ち」は，「どちらともいえない」が最も多く73人（47%），「とても幸福（うれしい）」と「少し幸福（うれしい）」を合わせて26人（17%），「少し不幸（悲しい）」と「とても不幸（悲しい）」を合わせても57人（37%）であった．

CATで難易度が調整されていることに気付くものは約半分で，問題が異なることを不公平と感じる者は少ない．ただし，ほとんどの者がCATは難しいと感じ，4割弱の者がCAT受験後に落ち込んでしまうようである．

Q5「CASEC(CAT)で何％ぐらいできたと思うか」，Q6 「高校の英語の得点」，Q7「テストで期待する得点」，Q8「テストで満足する得点」，Q9「テストでがっかりする点」について平均点をまとめると，図55のようになる．

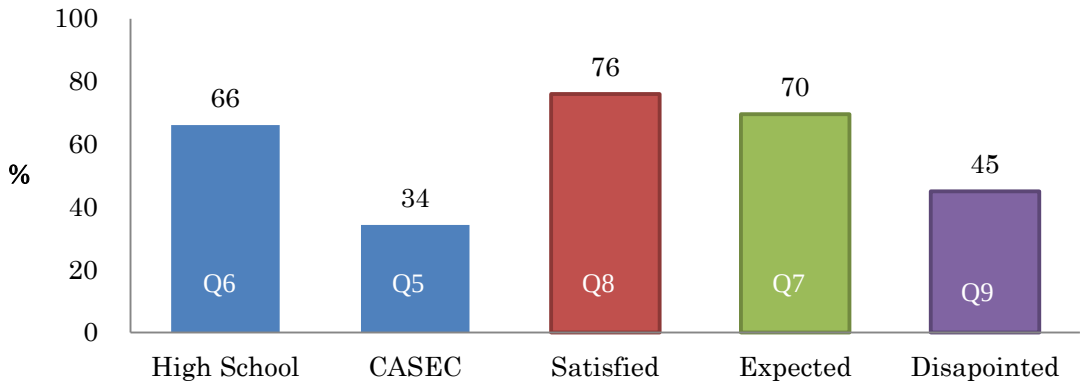

図 55　CAT 受験者のテストの得点についての意識

能力レベルに関係なく，ほとんどの者が CAT は難しいと感じ，4 割弱の者が CAT 受験後に落ち込んでしまうということは，現在ほとんどの CAT で採用されている情報量を最大化する項目

選択ルールでは，受験者の動機づけを低めたり，自己効力感を損なうことが懸念される．多少項目数が多くなっても，測定の精度が多少低くなっても，受験者の心理学的側面に配慮して，項目選択において目標正答確率を高くしてもよいのではないだろうか，理論編 2.3 でいくつかの先行研究を紹介したように，目標正答確率を上げることは受験者の自信を高めるが，あまり大きくないアイテムバンクで目標正答確率を上げると，精度が著しく落ちてしまうこともある．次節では，目標正答確率を Moodle-UCAT の機能を使って変化させて 2 とおりで CAT を実施し，目標正答確率の異なる 2 つの CAT に対する反応がどのように変わるかを調べて，CAT の心理学的側面について，もう少し考察を加えることにする．

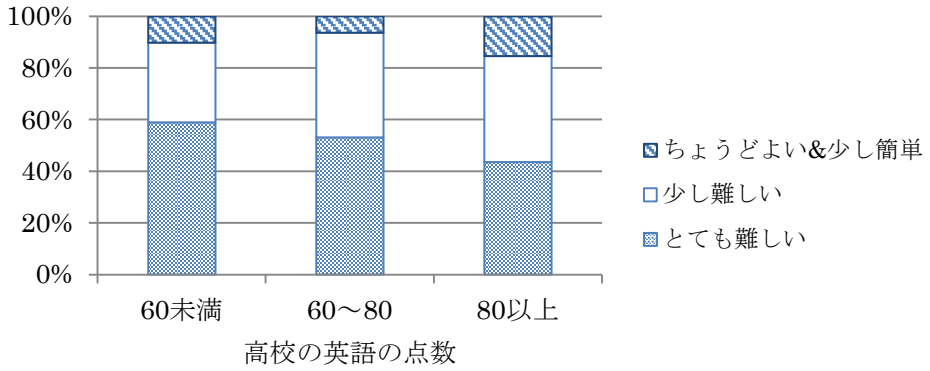

図 56　高校の英語の点数と Q1 への回答

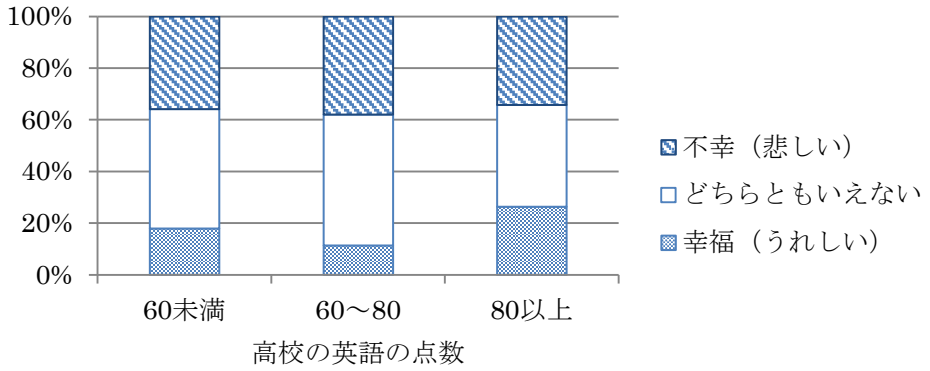

図 57　高校の英語の点数と Q4 への回答

9.2.　目標正答確率を変化させた場合の CAT に対する大学生の反応の変化

Kimura & Nagaoka (2012b)は，理論編 3.3 節で説明した Moodle UCAT を使って，CAT 実施時の項目選択ルールを調整し，通常の情報量を最大化にする目標正答確率を 50%にした CAT(A)と，易し目の項目が選択されるように目標正答確率を 70%にした CAT(B)を，約 200 人の大学 1 年生

に対して，1 か月の間を空けて実施し，それぞれの CAT 終了後に前節と同じアンケート調査を行った．2 つの CAT が同じ精度になるようにするために，表 36（Linacre, 2006）を参考に項目数を決定した．

表 36　同じ SE を得るために必要な項目数

Minimum number of CAT Items Administered					
Targeting	S.E. (logits)				
P	0.5	0.4	0.3	0.2	0.1
0.5	16	25	45	100	499
0.6	17	27	47	105	517
0.7	20	30	53	120	477
0.8	25	40	70	157	625
0.9	45	70	125	278	1112

各 CAT は Vgm のアイテム（7.6 節の表 25，8.3 節の図 51 参照）と Lng アイテム（7.8 節の表 31，8.3 節の図 51 参照）からなり，CAT はアイテムの種類ごとに実施された．2 つ CAT の目標正答確率，項目数，予想される SE をまとめると表 37 のようになる．

表 37　2 つの CAT の目標正答確率と実施項目数と予測される SE

		目標 正答確率	実施 項目数	予測され る SE
CAT(A)	Vgm	50%	16	0.5
	Lng	50%	16	0.5
CAT(B)	Vgm	70%	20	0.5
	Lng	70%	20	0.5

9.2.1.　CAT の結果

各 CAT の結果をアイテムバンクの種類ごとに示したものが表 38 である．いずれの CAT も 93~99%以上の受験者が予測どおり SE≦0.5 で終了していた．目標正答確率を 50%から 70%に高めても項目数を 16 から 20 に増やせば，CAT の結果は，理論どおり同じ精度で得られることが確認された．むしろ，目標正答確率を 70%に高め 20 項目実施した CAT(B)の方が，SE が 0.4 で終わるケースが 70~80%あり，平均値で見ても目標正答確率 50%で 16 項目実施した場合よりも，CAT で得られた結果の精度は高かった．ただし，まれに（全 CAT 受験者中 6 人）SE が 1.0 や 1.1 という非常に大きな値となっているケースがあったが，それは 1 問だけしか正解できていない受

験者や，1 問だけ不正解の受験者の場合であった．これは現状のアイテムバンクでは難しすぎる（あるいは易しすぎる）ケースである．

表 38　各 CAT 実施結果の基本統計量

	CAT(A)				CAT(B)			
	Vgm (N=240)		Lng (N=130)		Vgm (N=242)		Lng (N=215)	
	θ	*SE*	*θ*	*SE*	*θ*	*SE*	*θ*	*SE*
M	0.28	0.51	1.03	0.50	0.00	0.43	-0.14	0.44
SD	1.46	0.05	1.35	0.01	1.46	0.06	2.08	0.09
MAX	5.40	1.00	3.70	0.60	4.70	1.00	6.30	1.10
MIN	-3.70	0.50	-3.50	0.50	-4.50	0.40	-5.80	0.40

9.2.2.　アンケートの結果

Q1「難しさに」については，当然のことであるが，目標正答確率を上げた CAT(B)で「とても難しい」「少し難しい」ともに減少しており，「ちょうどうよい」という回答が増加している（図 58 参照）．全体の変化は χ^2 検定で統計的に有意な差（p=.008）があり，残差分析で「とても難しい」の減少と「ちょうどうよい」の増加は p<.01 の水準で統計的に有意な差が認められた．

Q2「難易度が調整されていることへの気づき」，Q3「問題が異なることへの不公平感」については，２つの CAT の間に，全体的に統計的に有意な差は認められなかった（図 59, 図 60 参照）．

Q4 の「受験後の気持ち」については，全体の変化は χ2 検定で統計的に有意な差（p=.014）があり，残差分析で「少し不幸（悲しい）」の減少は p<.05 の水準で，「少し幸福（うれしい）」の増加は p<.01 の水準で統計的に有意な差が認められた（図 61 参照）．

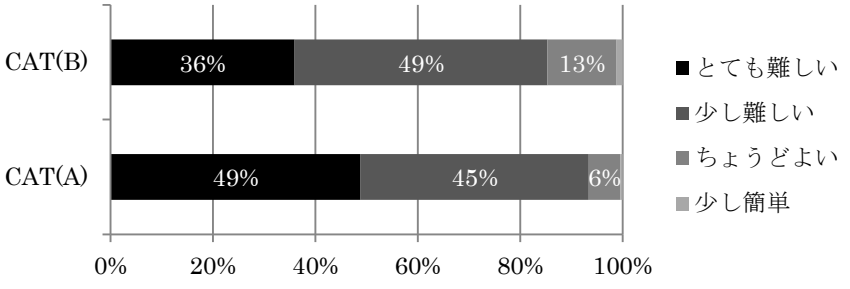

図 58　目標正答率の違いによる Q1 への回答への変化

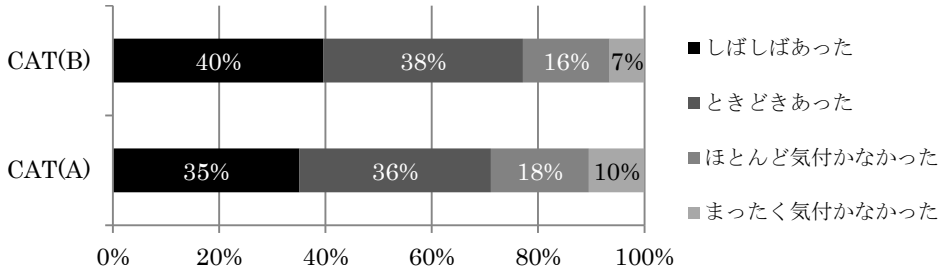

図 59　目標正答率の違いによる Q2 への回答への変化

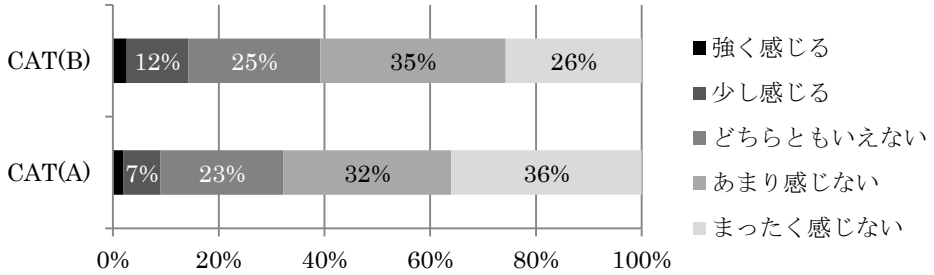

図 60　目標正答率の違いによる Q3 への回答への変化

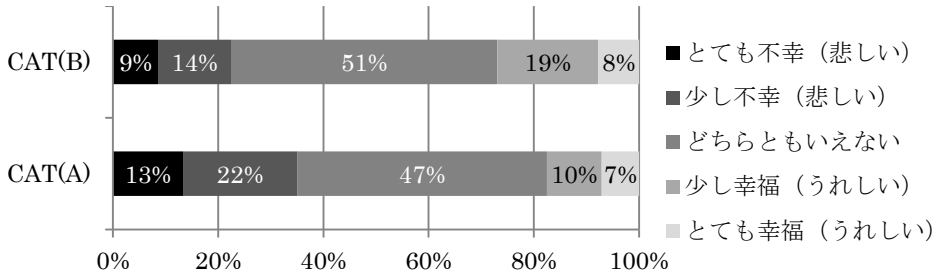

図 61　目標正答率の違いによる Q4 への回答への変化

Q5「CATで何％ぐらいできたと思うか」，Q6 「高校の英語の得点」，Q7「テストで期待する得点」，Q8「テストで満足する得点」，Q9「テストでがっかりする点」についての回答の平均点をまとめると，図62のようになる．CAT(A)受験後の回答とCAT(B)受験後の回答に加えて，図55（CASEC受験後のアンケート）の回答データも比較のために図62中に併記した．いずれも，ほぼ等質のグループに対して行ったアンケート調査なので，当然のことながらQ5以外の回答に差はみられなかった．Q5については，CASECについてもCAT(A)についても目標正答確率は50%であるのに，それぞれ34%と31%と低く，目標正答確率を70%にしたCAT(B)でも41%と低い．

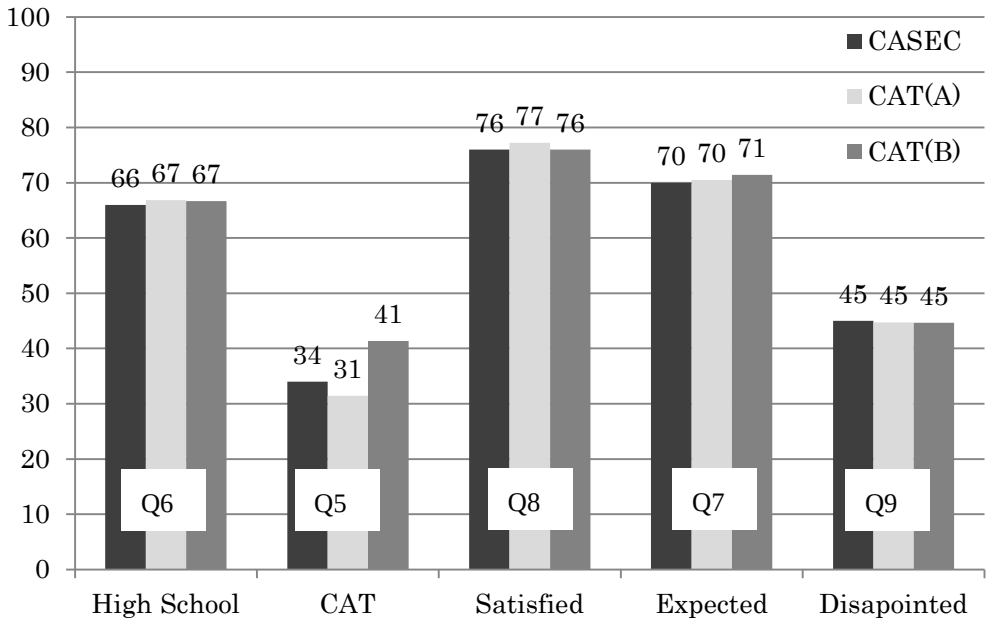

図 62　CAT 受験者のテストの得点についての意識の変化

目標正答確率に対して受験者が感じる正答率が低いのは, 受験者が過小評価しているのだろうか，あるいは受験者のレベルにあった項目がなくなり，計画通り目標正答確率を 70%にできなかったのであろうか，そのことを確かめるために，CAT(A)と CAT(B)の実際の正答率をアイテムバンクごとに調べたところ，平均値は CAT(A)の場合は Vgm で 43%，Lng で 48%と目標正答率 50%よりもやや低いがほぼ満たしている．CAT(B)の場合は Vgm で 58%，Lng で 56％と目標正答率より 10%以上低い結果になっている（表 39 参照）．ただし，分布のピークは，いずれの場合も，目標正答確率のところにある（図 63～図 66 参照）．CAT(A)Lng を除いて，分布はピークの左側の方が多くなっている．CAT(A)Lng だけ他と異なる傾向にあるのは，おそらく受験者数がこれだけ 130 と他より少ないためであろう. 分布のピークの左側の方が多くなっているということは，CAT の項目選択の段階で，平均よりも下のレベルの受験生に目標正答確率にそった項目が見つけられず，目標確率よりも少し正答確率の高い（少し難しい）項目を選択していたためではないだろうか. 第 8 章のシミュレーションと実テストによるアイテムバンクの検証でも，現在のアイテムバンクには, 平均よりも下のレベルの受験者の項目が不足していることが示唆されているので，十分考えられるシナリオである．さらに，考察を加えるためには，各受験者に対して選択された項目が, その時点の能力推定から判断される目標正答確率にどの程度近いものを選択できていたかを，細かく確かめる必要がある．ただし，現在の Moodle UCAT のモジュールから得られるデータから，それを計算することは，不可能ではないが大変効率が悪い．Moodle UCAT の CAT の記録の取り方に改良を加える必要がある．

表 39　各 CAT の受験者数と正答確率の平均値と標準偏差

	CAT(A)		CAT(B)	
	Vgm	Lng	Vgm	Lng
N	240	130	242	215
M	43%	48%	58%	56%
SD	13%	11%	14%	16%

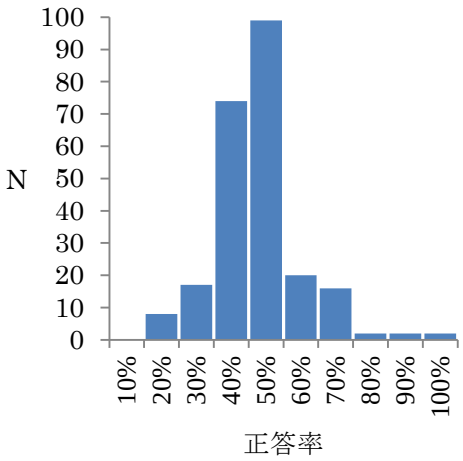

図 63　CAT(A) Vgm の正答率分布

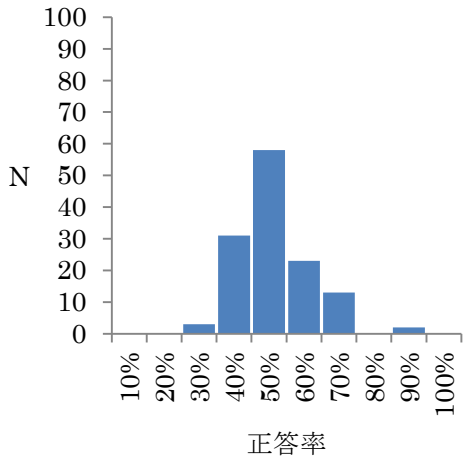

図 64　CAT(A) Lng の正答率分布

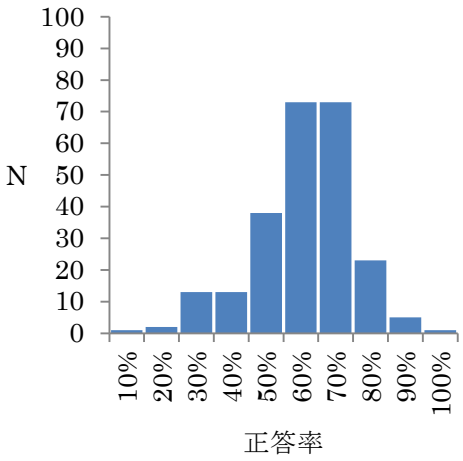

図 65　CAT(B) Vgm の正答率分布

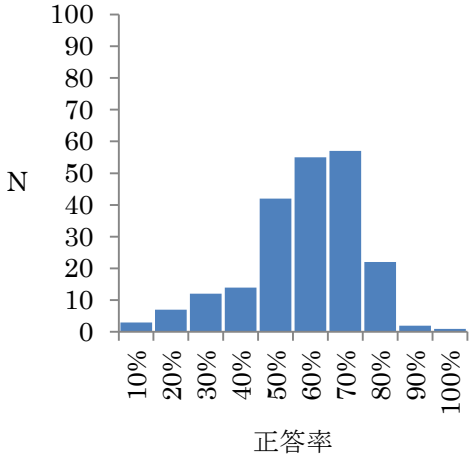

図 66　CAT(B) Lng の正答率分布

10. 診断的テスト結果の提示

テストは必ずしも学習評価のために行われるわけではなく, 入学試験のように選抜のために行われる試験もある. 資格試験もその試験で判定する能力がある一定水準以上か否かを判定するための試験であり, その試験で判定する能力を身につけるために行ってきた学習成果が試されるわけだが，その学習がどの程度成果を上げたか評価するものではない.

しかし，何らかの学習・教育の後に行われ，学習の成果を評価するテストの場合，どのようなことができるようになったのか, まだできるようになっていないことは何か, といった情報を提供することが望まれる．つまり，テストを実施した段階での，そのテストが測定しようとしている潜在能力について診断的情報を提供できることが望ましい．しかし, 教育現場で実施されているものでそういった診断的情報提供を目指したテストはほとんどなく, そのテストを受験した集団の中で，相対的にどのような位置にあるかによって評価されるか，テストの点数そのもの（何パーセント正解したかなど）で評価されることがほとんどである.

テストが学習した能力を測定しているという概念的妥当性が満たされているとして, テストとそれに基づく学習評価の関係は, 前者が診断的情報を提供することで後者の説明内容に具体性が生まれるようになることが健全であり, そのような関係が築けるように教育者とテスト開発者は努力すべきである.

以下，本章では LRT の RMP を使った診断的テスト結果の提示方法の可能性を探るとともに, CDS を使った自己評価について LRT で分析した結果について考察を加え, さらに CDS を使った自己評価とテスト結果の比較を行い，テスト結果を CDS に結びつける可能性について探る.

10.1. LRT による診断的テスト結果の提示

LRTの受験者能力のとらえ方について1.5.2で説明したように，LRTでは，受験者の潜在ランクを順序尺度上に推定することだけでなく，受験者が各ランクに所属する確率を集めたRMPとして示すことで、結果を多義的に表現することができる. 推定された潜在ランクが同じだったとしても，RMPの形状の違いによって，受験者に異なるフィードバックを返すことができる．また，同一の学習者のRMPを時系列で追いかけることで，学力の変化を多義的に表現することも可能である.

ここでは，RMPの変化としてどのようなパタンがあるかを探るとともに，RMPを提示することで, テスト結果にどのような診断的情報を加えることが可能かを探る. 4月と8月に実施された26項目からなるVgmの2つのテスト2Aと2Gの両方を受験した大学1年生70名のデータから，大きく分けて4つのパタンがあることがわかった．なお，両テストは事前の分析（木村, 2009a）により項目特性を調べて固定されたアンカー項目を6項目ずつ含み，等化が図られている（木村・永岡, 2010b）.

（1） 推定ランクとして変化はないが下位ランクへの所属確率が減少し，上位ランクへの所属確率が増加しているケース

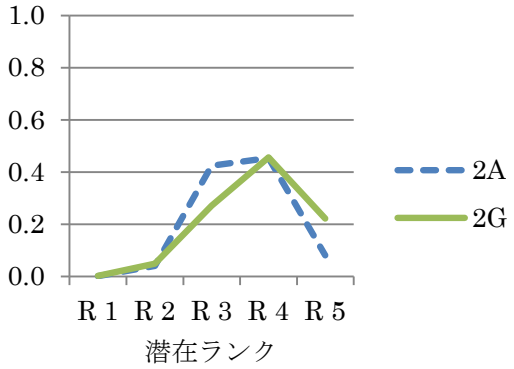

図 67 RMP の変化例（1－1）

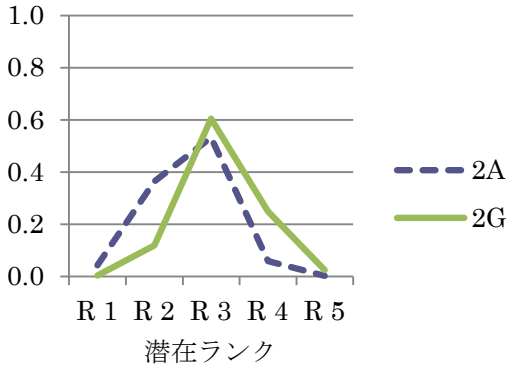

図 68 RMP の変化例（1－2）

このタイプの場合は，２つのテスト結果を潜在ランクだけで示すと「変化なし」となるが，RMP の変化を示すことで，前よりも上のランクに上がりつつあることを受験者に伝えることができる．

（2） 推定ランクとして変化はないが下位ランクへの所属確率が増加し，上位ランクへの所属確率が減少しているケース

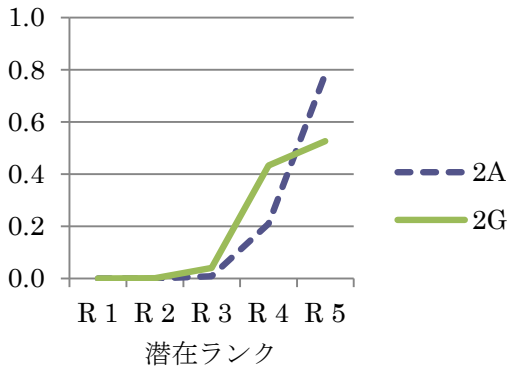

図 69 RMP の変化例（2－1）

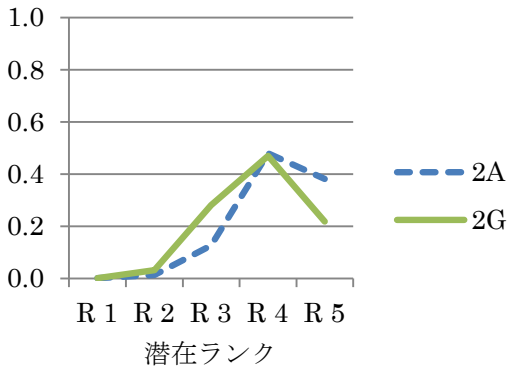

図 70 RMP の変化例（2－2）

このタイプの場合も，２つのテスト結果を潜在ランクだけで示すと「変化なし」となるが，RMP の変化を示すことで，前よりも下のランクに下がりつつあることを受験者に伝えることができる．

（３）　推定ランクが上昇：所属確率の変化が少ない場合・多い場合

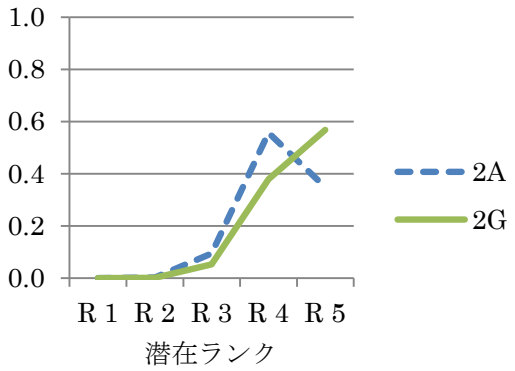

図 71　RMP の変化例（３－１）

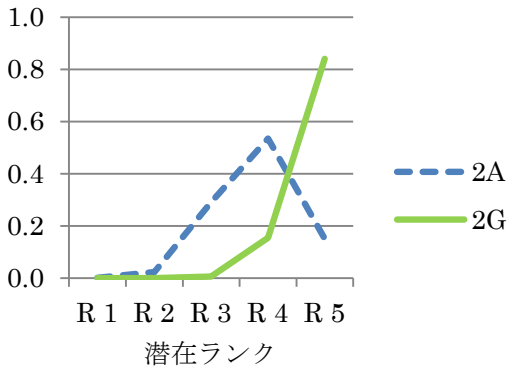

図 72　RMP の変化例（３－２）

このタイプの場合は，２つのテスト結果を潜在ランクだけで示すと「ひとつ上のランクに上がった」となるが，RMP の変化を示すことで，そのことの確からしさを受験者に伝えることができる．左の図 71 の場合よりも，右の図 72 の方が，より確実にひとつ上のランクに上がったことが示せている．

（４）　推定ランクが下降：所属確率の変化が少ない場合・多い場合

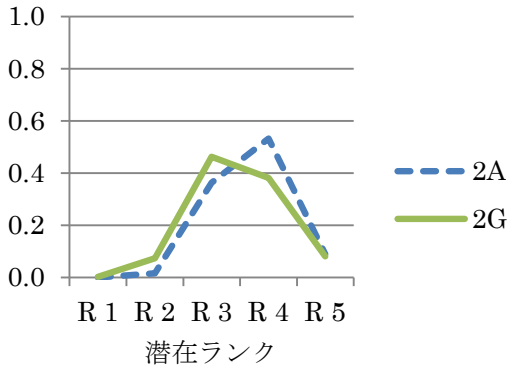

図 73　RMP の変化例（４－１）

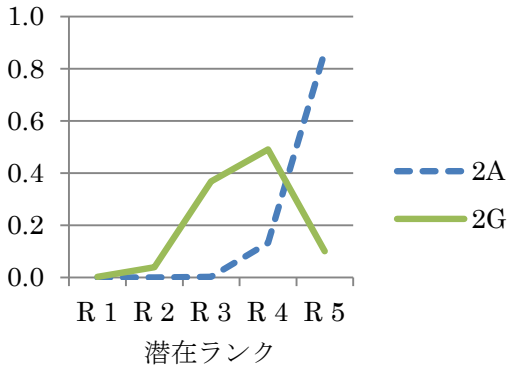

図 74　RMP の変化例（４－２）

このタイプの場合は，２つのテスト結果を潜在ランクだけで示すと「ひとつ下のランクに下がった」となるが，RMP の変化を示すことで，そのことの確からしさを受験者に伝えることができる．左の図 73 の場合よりも，右の図 74 の方が，より確実にひとつ下のランクに下がったことを示せている．

10.2. CDS による診断的情報の提供と自己評価

テストに基づく評価に対する考え方は, そのテストが測定しようとしている潜在能力について, テストを実施した母集団の中で, そのテストを受験した個人あるいは集団がどの位置に属するのかを判断することができればよいという考え方が, 観点別評価の導入(平成元年度学習指導要領)までは大勢を占めていた. 学校教育では平成元年の学習指導要領改定を期に, 観点別に絶対評価や到達度評価が行われることになったが, それによって期待されたほど学習改善指導が向上したとは言えず, 標準学力テストやそれに基づく受験指導などではまだ集団基準の相対評価が根強く残っている.

学習者と教育者が, それまでの学習の振り返りとその後の学習・教育計画のために, テストの結果から必要としている情報は, 集団基準の相対評価でも大まかな観点別評価でもない. テストが測定しようとしている潜在能力を使って, 個々の学習者が現段階で「どんなことができるのか／できないのか」といったより具体的な診断的情報である. 実施したテストの結果とCDSを結びつけることが, 学習者にとっても教育者にとっても, 重要な情報を与えることにつながると考えられる.

2001 年に正式公開された「ヨーロッパ言語共通参照枠（Common European Framework of Reference for Languages: CEFR)」は, 欧州評議会（Council of Europe）が, EU における外国語教育向上, 第 2 言語の使用, 教育方針や学習達成度などで共通理解を可能にする目的で開発したもので, これに準拠したテストや学習評価では, 外国語能力を 6 つのレベルに分けて認定する (Council of Europe, 2001). 各レベルは総合的にも能力下位区分ごとにも descriptor と呼ばれる CDS が細かく提供されており, それを参照することで, 各レベルの学習者が, 具体的にどのようなことができるかを把握できる.

CEFR に触発されて, 日本国内での CEFR の適用や, 日本版 CEFR の開発などがいくつも手がけられるようになった（斉田, 2008; 投野, 2010; 根岸, 2011）. また, 日本英語検定協会（The Society for Testing English Proficiency, STEP）も CEFR を参考に大規模な調査を行い, 英検の各級と CEFR のレベルの対応づけ（Dunlea, 2009; Dunlea & Figueras, 2010）や, 英検各級に対して 4 技能別に CDS のリストを公開している. これらの研究・開発により, テスト結果を受験者集団の中の相対的位置で評価するのではなく, 診断的情報を提供する素地が生まれてきている. しかし, まだ日本での共通参照枠として全体で共有できるものが完成しているとはいえない.

2007 年に制定された教育基本法（平成１８年法律第１２０号）において,「自ら進んで学習に取り組む意欲を高めることを重視して行われなければならない.」と明記してあり, 学校教育における主体的に学習に取り組む態度を育成することが示されている. 主体的に学習に取り組む態度を育成するには「学習に対するメタ認知」が重要であることは, 竹内(2007)など多くの教育者・研究者が主張するところである.

大学英語教育学会学習ストラテジー研究会（2005）は, 英語力を育成するのには, 学習者の自律を重視し, 学習者が効果的に自らの力で英語学習をできるような学習ストラテジーを立てさせることが重要であると主張している. 同研究会は, 主要な学習ストラテジーを「メタ認知ストラ

テジー」「認知ストラテジー」「社会・情意ストラテジー」に３分類し,「メタ認知ストラテジー」の中には「理解度をチェックする（evaluate yourself)」すなわち自己評価を入れている.

自己評価をテストの一部に組み込んだコンピュータを利用したテストとしては，EUで外国語能力を診断的に判定するために開発されたDIALANG[21]がある．これはオンラインでもダウンロードしても受験できるもので，14 の言語について，語彙・読解・聴解・作文・文法の 5 つの領域のテスト用意されており，最初に語彙レベルを診断すると，その結果に基づき他の領域は３つの異なるレベルのテストが実施される. すべての領域を受験せずに必要なものだけを選択して受験することもできる．この５つのテストの他に，CEFRに準拠したCDSに対する自己評価も求められる．各領域のテスト結果はCEFRの 6 つのレベルで示されるだけでなく，自己評価とのずれ（過大評価あるいは過小評価）についてもフィードバックされる（Alderson 2005）.

このようにテスト受験者に自己評価させることは, 学習の成果を自ら振り返らせることになり, その後の学習計画や目標を立てる上でも役立つ. テストの評価とのずれを認識することで, 過大評価あるいは過小評価を修正することもできる.

学習改善と指導計画改善につながる学習評価を行うには，1）テストの結果をなんらかの CDS と結び付けて呈示することと，2）テストと平行して自己評価を学習者に行わせ，学習を振り返り自分の能力の変化への気づきを促すこと，3）自己評価とテストの評価のずれ（自分の能力に対する過大評価や過小評価）に気づかせ修正させることが望まれる．また，テストおよび自己評価の実施・採点・集計にコンピュータを利用することで，作業を容易にするとともに，結果フィードバックの即時性を高めることなどが期待される.

10.3. 英語教育における CDS を使った自己評価

新潟県内の 2 大学の 1 年生対象に開講された一般教養科目の英語 7 クラスの受講生合計 295 名（工学系・看護系・福祉心理系の学部学科）の協力を得て，STEP の英検 Can-do リストの CDS を使って，入学時の英語力を自己評価してもらった．分析には，4 技能すべての自己評価を行った 233 名のデータを対象に始めたが，途中ですべてに Yes または No とするなど，明らかに誠実に回答していないと思われる者を分析対象から外したので, 分析した回答者数は Listening が 217, それ以外の 3 技能が 220 となった.

STEP の CDS は 2003～2006 年に 1 級から 5 級の合格者のべ 20,000 人超に対してアンケート調査を実施し, その結果によって作成されたものである. 本研究ではこのうち英検 4 級～準 1 級の 4 技能（R: reading, L: listening, S: speaking, W: writing）別の CDS を学習者の自己評価のツールとして用いることとした（表 40 参照）. STEP の調査では，共通項目を配置しながら各級ごとに質問紙が作成され，各級の合格者に対して，自信の度合いを 5 段階で回答する形式であったが，本研究では協力者全員に, 表 2 に示した数の CDS を技能ごとにすべて Yes/No で回答するように求めた. 5 段階でなく Yes/No で回答を求めたのは，協力者に多くの CDS を見て判断してもらうこ

[21] http://www.dialang.org/

とを優先したためで, 回答時間のよりかからない方法としてこの方法を選択した. 回答データは Moodle のフィードバックモジュールで収集し，Excel 形式でエクスポートしたデータを Exametrika により LRT に基づいた分析を行った. すでに, RM に基づいた分析は Kimura & Nagaoka (2011a) で WINSTEPS を使って行っているが，ここでは，LRT に基づいた分析結果を中心に述べ，考察を加える.

表 40　利用した英検 Can-do リストの CDS の数

	R	L	S	W
準1級	6	4	4	4
2級	5	5	7	6
準2級	4	6	6	5
3級	6	6	6	5
4級	5	6	7	6
合　計	26	27	30	26

CDS の困難度の順序性が保たれているかどうか検証するために，4 技能ごとに各 CDS の IRP 指標 β と，その CDS が所属する級の順位相関を調べた．英検 4 級を 1 に，3 級，準 2 級，2 級，準 1 級をそれぞれ 2，3，4，5 と序数化して，β との間の相関がどの程度あるか，スピアマンの順位相関を求めた．全般的に β の方が小さくなるケースが多い．これは協力者の英語力のレンジがほぼ英検 3 級から 2 級に収まっているため, 英検 4 級と 3 級の CDS の両方に対して「できる」と回答した者が多かったため，この２つの級の CDS がどちらも同じランク 1 に分類されることが多かったためであろう.

また, 完全に一致している割合は R と W が 72%と 62%と比較的高いのに対して, L と S が 44%と 47%と低いのは，協力者たちが実際に英語を聞いたり話したりする機会は，英語を読み書きする機会に比べて極めて少ないためだと推察する．つまり，L と S について比較的難しい内容の CDS を見ても，文字媒体しか使っていない状況で，経験したことのない音声活動のことを想像しても,「(やったことはないが) これくらいは，もしやればできるだろう」と判断してしまうのではないだろうか.

表 41 に示すようにいずれの技能についても.93～.95 と非常に高い相関を示した．本研究で分析した CDS の困難度は，STEP の調査結果とその順序性がほぼ一致していると言ってよいだろう．RM に基づいた分析は Kimura & Nagaoka (2011a) でも，4 技能の項目の信頼性は，いずれも 0.99 と非常に高かった．しかしながら，CDS の所属英検級と IRP 指標 β がすべて一致しているわけではない（表 42 参照）.

全般的に β の方が小さくなるケースが多い. これは協力者の英語力のレンジがほぼ英検 3 級から 2 級に収まっているため，英検 4 級と 3 級の CDS の両方に対して「できる」と回答した者が多かったため，この２つの級の CDS がどちらも同じランク 1 に分類されることが多かったため

であろう.

また, 完全に一致している割合は R と W が 72%と 62%と比較的高いのに対して, L と S が 44%と 47%と低いのは，協力者たちが実際に英語を聞いたり話したりする機会は，英語を読み書きする機会に比べて極めて少ないためだと推察する．つまり，L と S について比較的難しい内容の CDS を見ても，文字媒体しか使っていない状況で，経験したことのない音声活動のことを想像しても,「(やったことはないが) これくらいは，もしやればできるだろう」と判断してしまうのではないだろうか.

表 41　CDS の所属英検級と IRP 指標 β の順位相関

R	L	S	W
0.93	0.94	0.94	0.95

表 42　CDS の所属英検級と IRP 指標 β の一致度

級－β	R	L	S	W
-2	0	0	0	0
-1	4	1	1	2
±0	20	14	18	17
+1	1	11	10	7
+2	1	1	1	0
総数	26	27	30	26
一致%	77%	44%	47%	62%

英検Can-doリストのCDS全体の順序性を乱していると考えられるのは，表42の中の「級－β」が+2となっている3つのCDSである．これらのCDSへの回答に対して今後も同じ傾向が表れるようなら，何らかの修正を加えるかリストから除外した方がよいかもしれない．以下に，そのCDSを示しながら原因を考察する.

(1)「簡単なチラシやパンフレットを理解することができる．(商品の値段，セールの情報など)」(英検2級のRのCDS，β＝2)：同じ2級のRのCDSが「一般向けに書かれた実用書」や「日本語の注のついた英字新聞」を題材としたものなので,「チラシやパンフレット」は相対的に易しいと感じたのかもしれない．しかし，実際に英語の「チラシやパンフレット」を手にした経験がない者が多いと予測されることも，この結果に影響を与えているかもしれない.

(2)「簡単な道案内を聞いて，理解することができる．　(例：Go straight and turn left at the next corner.)」(英検準2級のLのCDS，β＝1)：道案内は初級英会話でよく扱うトピックであることと，実際は音声を聞いて理解できるかが問題なのだが，文字で提示されていることで，より易しいと感じたのではないだろうか.

(3) 「自分の気持ちを表現することができる．（うれしい，悲しい，さびしいなど)」（英検準2級のSのCDS，β＝1）：文の意味するところが少し曖昧に思える．感情が表現できるということは，感情を表す単語と表現（I'm happy.　I feel sad.など）を知っていることだけなのか，単語レベルだけでなく抑揚など音声レベルに変化をつけて感情を表現することも含むのか．後者は比較的難しいが，多くの回答者は前者の単語レベルのことだけを想定して回答したのではないだろうか．

10.4. 英語教育における CDS を使った自己評価とテスト結果の比較

前節のCDSを使った自己評価への協力者のうち，リーディング（Rdg）とリスニング（Lng） のテストを受験したものについて，自己評価とこれらのテストの結果にどのようなずれが生じているか比較した（木村，2012）．これを技能ごとにまとめたものが表43と表44である．

表 43　リーディングの自己評価とテスト結果のずれ

		テスト結果（ランク）					
		R1	R2	R3	R4	R5	合計
自己評価（ランク）	R1	12	2	3	8	7	32
	R2	6	3	5	3	7	24
	R3	4	1	4	2	6	17
	R4	3	3	5	3	8	22
	R5	4	1	1	2	13	21
	合計	29	10	18	18	41	116

表 44　リスニングの自己評価とテスト結果のずれ

		テスト結果（ランク）					
		R1	R2	R3	R4	R5	合計
自己評価（ランク）	R1	10	10	6	3	2	31
	R2	9	8	2	1	5	25
	R3	6	7	3	3	7	26
	R4	2	6	2	4	7	21
	R5	6	3	7	5	15	36
	合計	33	34	20	16	36	139

自己評価とテストの結果のランク数が完全に一致しているのは，リーディングで35人（30%），

リスニングで40人（29%）である．しかし，実際に行ったことがない事象についての記述も多いCDSに対して，必ずYes/Noで回答するのが難しいことも考慮に入れると，自己評価とテストの結果が少しだけずれることは自然なことと考えられるので，自己評価とテストの結果のランク数が1だけ上下にずれている場合も，ほぼ一致しているとみなすこととした．その基準で表43と表44をもう一度見ると，リーディングで66人（57%），リスニングで85人（61%）が自己評価とテストの結果が一致しており，約6割の学習者は自己評価とテストの結果にずれがほぼないことがわかる．

自己評価のランク数がテストの結果のランク数より2以上大きい場合を，過大評価（overestimate），自己評価のランク数がテストの結果のランク数より2以上小さい場合を過小評価（underestimate）と定義することにした．今回はランク数を5とした分析なので，いずれの場合も最大4段階の差が生じうるが，いずれも１～4%と極めて少ない．

リーディングとリスニングを比べると，過小評価となる者は，34人（29%）対24人（17%）でリーディングの方が多く，過大評価となる者は，16人（13%）対30人（22%）でリスニングの方が多い（図75と図76参照）．この割合の差は統計的にも有意な差である（$\chi^2 = 5.84$, $df = 1$, $p = .016$）．これは4.1で考察したように，リスニングの方が，CDSに書かれている内容を実際に経験する機会が少ない（あるはまったくない）ことと，音声を伴わない文字媒体だけの状況での回答なので，「（やったことはないが）これくらいは，もしやればできるだろう」と判断してしまったためであろうことが推察される．

テストを実施するだけでなく，CDSを使って自己評価をさせることは，受験者が自分の能力を過大評価（あるいは過小評価）していることを気づかせるきっかけとすることもできるであろう．テストによる能力評価とCDSによる自己評価は，基本的に別次元のことを測定しているので，単純に両者をむすびつけることはできない．しかし，一般的にテストの結果とCDSを結びるけることも，検討する価値が十分ある．

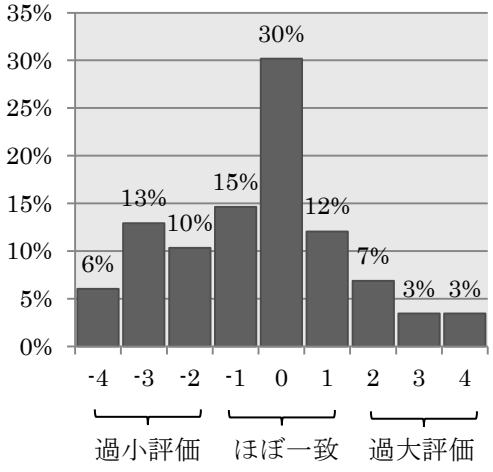

図 75　過大評価と過小評価の割合（Rdg）

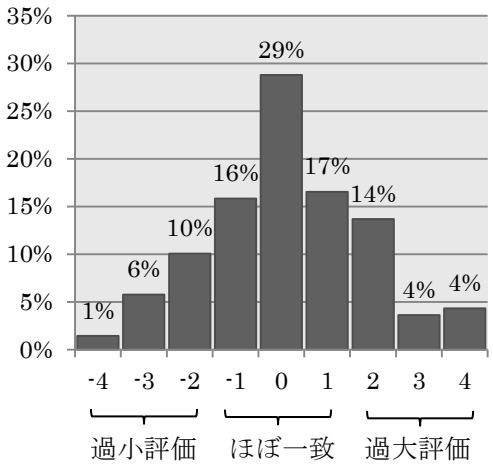

図 76　過大評価と過小評価の割合（Lng）

まとめと今後の課題

本研究の目的は，理論と実践の両面からコンピュータ適応型テストについて検討を加え，新規のアルゴリズムを提案するとともに，オープンソースを利用してそれを実装するシステムを開発し，そのシステムを英語教育することによって検証することであった．その中で，理論的な側面として，潜在ランクに基づくコンピュータ適応型テスト（LRT-CAT）に関して次の 2 つの提案を行った．

（１） CAT を実装するためのアイテムバンクを構築する際に，望ましくない項目を LRT の IRP 指標を使って除去する方針

モデル適合度の指標は既に提案されていたが，CAT 実装するためのアイテムバンクの項目として望ましくない項目を，LRT の枠組み（IRP 指標）を使ってヒューリスティックな判断で取り除く指針を提案した．ただし，提案した基準は絶対的なものではなく，今後より多くのデータにあたり検討を加える必要がある．他のモデルで除去される項目と一致しない部分について，さらに考察を加えることで，LRT の特徴をさらに明らかにすることもできるであろう．

（２） LRT-CAT アルゴリズム：項目選択ルールと終了条件

LRTは段階評価に適したテスト理論であり，教育現場でのその有用性は高いが，新しいテスト理論であるため，これまでCATのアルゴリズムについての提案はなされていなかった．本研究は，LRTの特徴であるRMPとIRPに焦点を当てて項目を選択するルールと，RMPの変化量に着目して終了条件を立てる案を提案した．

項目選択ルールとしては，(37)式により求められる λ によって暫定RMPに対する各項目の識別度の高さを評価し，CAT終盤に識別度の高い項目を温存し，CAT初期では識別度の低い項目から選択されるようにするために，λ が最小である項目から出題することを提案した．終了条件としては，RMPの変化が一定以下になった場合に，RMPの推定が収束したと判断して終了させる方法として，(38)式の μ の値を使うことを提案した．

今後，他のルール（たとえば，λ が最大である項目から出題した場合など）で項目を選択した場合と比較検討することも大変興味深い研究テーマとなるが，その際にはCATに利用するアイテムバンクの特性（サイズと集まった項目の分布状況）とともに考察する必要があるだろう．サーバーの計算負荷を軽減するためにIRP指標βが受験者の暫定の潜在ランクの推定値±1の項目に限定して λ の値を計算したが，この制限を加えずに選択した場合にどの程度の計算負荷がかかるのか，選択結果にどのような違いがあるかについて，検証することも興味深い．また，目標正答確率を項目選択ルールに取り入れた場合，どのような結果になるかも今後の課題である．LRT-CATにおいて，正答確率（困難度）を項目選択ルールに導入するならば，(43)式によって求められるRMPとIRPの積和（受験者iの項目jに対する正答確率）

を使うことができる.

大学入学生の英語基礎力を測定する小規模な CAT を開発するために行った一連の実践的研究で，次の（3）～（7）の 5 つのことを示すことができた.

（3） CAT を実装するためのアイテムバンク構築は， 小規模であっても，RM または LR T に基づき行えること

RM では 100～200 程度のサンプルサイズで十分項目分析できることが既にわかっているたが，LRT でも分析する潜在ランク数を 5 程度に少なくすることで，十分項目分析できることを実践的研究で示した.

（4） RM に基づく項目分析も，LRT に基づく項目分析も，オープンソースを利用して十分行えること

RM においても LRT においても，小問形式の 2 値データと大問形式の多値データの両方を，オープンソースを利用して十分に分析可能であることを実践的研究で示した.

（5） CAT を実装するシステムをオープンソース LMS である Moodle を使って開発できること

RM に基づく BASIC プログラム UCAT を，Moodle で実装できるようにプログラムを書き換え利用できるようにした．上記（2）の提案に基づき Moodle 上で CAT を実装するモジュールも開発された．これらのモジュールについては，イギリスに拠点を置く世界規模の語学学校 Kaplan Interantional Colleges，シンガポールで e ラーニングを学校・企業に提供する企業 ACP などから照会があり，今後のシステム機能向上のために共同研究を行う予定である．開発した現モジュールについて英文ドキュメントを完成させ，研究成果を広く公開する予定である.

（6） シミュレーションにより，用意できたアイテムバンクを使って CAT を実装した場合，どの程度の結果が得られるか，アイテムバンクのどこに弱点があるかを把握できること

LRT-CAT のシミュレーションを既存のアイテムバンクを使って行い，実際に CAT を実装し，指定項目数で終了させる場合に，何項目にすべきかを検討するとともに，既存のアイテムバンクのどこに弱点があるかを分析した.

今回はシミュレーションにおける終了条件として，暫定RMPの変化に着目し，(38)式によりμの値が0.05未満になった場合に暫定RMPの変化が十分小さくなったと考え，現アイテムバンクでCATを指定項目数で終了させる場合に何項目にすべきかを探った．μの値を変化さ

せて，さらにシミュレーションを行えば，より適切な終了条件を見極める材料が得られるかもしれない. ただし, 現アイテムバンクには, 潜在ランクが2の受験者に出題すべき項目（IRP指標β=2である項目）が少ないので，このレベルの項目をアイテムバンクに追加してから，再検証するべきであろう.

（７） 用意できたアイテムバンクを使って実際に CAT を実施し，どのレベルの項目の使用頻度が高くなるか（どのレベルの項目を今後追加すべきか）を把握できること

RM-CAT を Moodle UCAT を使って実際に実施し，項目の使用頻度を調べることで，どのレベルの項目が頻繁に使われているか（どのレベルの項目が不足しているか）を把握した.

また，CAT 受験者に対するアンケート調査からは，次の（８）～（１０）の３つことを明らかにすることができた.

（８） 先行研究で指摘されていた通り，通常の CAT の項目選択ルールでは，ほとんどの受験者がテストを難しいと感じ，多くの者が受験後に落ち込んでしまうこと

測定の精度を優先すれば，情報量が最大になる項目（目標正答確率 50%）を出題するのが最も望ましいが，受験者の心理的側面を考慮するならば，情報量を多少犠牲にして，実施項目数を増やしても，目標正答確率のもう少し高い項目が選択されるようにすべきである.

（９） 目標正答確率を通常の CAT の項目選択ルールの 50%から 70%に変更すると，難しいと感じる受験者は少なくなり，受験後に落ち込む者も少なくなること

同じアイテムバンクを使って, 同じ受験者集団に目標正答確率 50%と 70%の 2 種類の CAT を実施し，受験後アンケート調査をしたところ，当然のことながら，目標正答確率 70%の CAT 受験後の方が，難しいと感じる受験者は少なくなり，受験後に落ち込む者も少なくなった. ただし, いずれの場合もどの程度正答できたかという印象は設定されている目標正答確率を大きく下回った. その原因については，今後システムの改良と，さらなる実践的研究が必要である.

（１０） 目標正答確率を通常の CAT の項目選択ルールの 50%から 70%に変更しても, 実施項目数を 16 から 20 に増やすことで，同じかそれ以上の精度で CAT が終了すること

理論的にほぼ同じSEで終わると考えられる，目標正答確率50%で16項目のCATと，目標正答確率70%で20項目のCATを実施し，理論どおり同じSE以下で終了することを確認した. むしろ，目標正答確率70%で20項目のCATの方が高い精度で終わっているケースが多かった. 本研究ではCAT全体をとおして目標正答確率を変更したが，CATのどの部分の目標正答確率

を上げるのかによっても，受験者の印象は異なる．先行研究でも，すべてでなく一部の項目を80%にする方が，全体を70%にするよりも測定精度に問題が起こらないことが指摘されている．今後は，目標正答確率を部分的に調整できるアルゴリズムを開発して，受験者の印象をさらに調べることで，重要な知見がえられるであろう．

診断的情報を CAT の結果に付加する方策については，LRT における受験者の潜在能力のとらえ方である RMP を利用することと，CDS を使った自己評価の試みから，次の（１１）（１２）の２つのことが有効であることを示した．

（１１） LRT に基づき分析した場合，潜在ランクだけでなく RMP を示すことで，変化の状況をより細かく示すことができること

LRT の場合，潜在ランクを示すだけであると，その情報が少なく変化を示しづらいが，RMP を示すことで，潜在ランクが変化していなくても，上昇傾向にあるのか下降傾向にあるのかを示すことが可能で，受験者が結果を解釈する際や，教員（テスト実施者）が結果についてフィードバックを与える際に参考になる．

（１２） CDS を使って自己評価することは，自分の能力を過大評価（あるいは過小評価）していることを気づかせることができること

CDS に対する評価は先行研究のものとほぼ同じ順序性があることが確認された．テストによる評価だけでなく，CDS を使った自己評価を併用することは，自分の能力を過大評価（あるいは過小評価）していることを気づかせる上で有効な方法であろう．しかし，実際にテストで測定している潜在能力と，CDS に表されていることができるかどうかは，別次元のことであり，単純に両者を結び付けることは難しい．

今後の課題として，まず取り組まなければいけないことは，読解力問題（Rdg）の CAT を実装するシステムを開発することである．Rdg の項目は大問形式であるため，多値型アイテムとしてアイテムバンクは構築されているが，それを使って CAT を実装するシステムがまだない．RM の多値型アイテム用アルゴリズムとしては Lnage (2007) がシミュレーションによりその方向を示しており参考になる．実践例としてはデンマークの初等・中等教育で大規模に行われている事例（Albeck，2012）が参考になる．LRT の多値型アイテム用アルゴリズムについては，LRT の２値型アイテム用アルゴリズム（木村・永岡，2012a）をもとに提案を行い，それを実装するシステムの開発を行いたい．これまでは，Moodle 上に CAT を実装するための追加モジュールを開発するアプローチをとってきたが，それでは Moodle の基幹プログラムがバージョンアップされるたびに，開発したモジュールを修正する必要が生じる．今後は，CAT の実装自体はオープンソースの CAT 実装システムである Concerto 上で開発し，LMS である Moodle との間でデータ連

携を図るアプローチが有効かもしれない.

本研究では取り上げることができなかった CAT の問題はいくつもある．その中でも，最大のものは，コンテントバランスをどのように保つかという問題であろう．これについては，1 問ごとに項目を選択する方式では，van der Linder (2010)の提案する shadow tests という方式が有効であると考えられるが，CAT のアルゴリズムは複雑になる．一方，1 問ごとに項目を選択するのではなく，コンテントバランスの調整をした，難易度の異なるテストを複数用意して，テストを 2 や 3 のステージに分けて実施する multi-stage test (MST; von Davier & Haberman, 2012)という方法は，コンテントバランスの調整が比較的容易で，今後ますます注目を集めると思われる．MST は，CAT と異なり 1 つのステージの中のテスト項目は固定なので，項目をどの順番で解いてもよく，一度解答した問題を後で見直して変更することもできるという利点もある．すでに，アメリカの大学院出願者が受験するテスト GRE の Verbal Sections と Math（Quantitative）Sections は，2007 年から，1 問ごとに項目を選択する CAT から MST に変更されている．

その他にも，コンピュータの特性を生かした新しいタイプの項目開発は，今後の大きな課題であろう．これまでの CBT や CAT の項目は PPT の項目をコンピュータ化しただけのものが大半である. マルチメディア性やインタラクティブ性といったコンピュータの特性を生かした項目が開発されると, PPT では測定が困難であったコミュニケーション能力や問題解決能力などを測定しやすくなるのではないだろうか. マルチメディア性やインタラクティブ性をもった新しい項目がテストに加わることで，CBT や CAT に対するイメージが変わり，受験者・学習者へのポジティブな wash back 効果が生じることが期待される．また，CAT にゲーム的要素を入れること（gamification）によって，学習への動機づけが強化されることも期待される．いやいやテストを受けるのではなく，CAT にやみつきになり解答を続けているうちに，気がつくとスキルや知識が身につくようになるかもしれない.

本研究によって，段階評価に適した新しいテスト理論である LRT-CAT アルゴリズムが提案され，オープンソースを使った CAT 開発の道筋を示せたことで，一人でも多くの方が CAT というツールを使って教育測定をできるようになれば幸いである．規模の大きさにかかわらず， CAT を自分の手で開発し実行できる時代になったのである. これから CAT を開発しようとする方は，ぜひ受験者の心理的側面にも配慮した CAT を目指していただきた.“Happy CAT for everyone”これがこれまで筆者の研究モットーであり，これからも変わることはない.

謝辞

本研究を完成させるために，ご指導・ご助力いただいた指導教員の永岡慶三教授に深く感謝の意を表します．副査の先生方（野嶋栄一郎教授，森田裕介教授，ならびに大学入試センターの荘島宏二郎准教授）から，多くの示唆をいただけたことに衷心より感謝を申し上げます．

LRT に関しては荘島宏二郎准教授に，RM に関しては Mike Linacr 元 Sydney University 教授と James Cook 大学の Trevor Bond 教授に，CAT に関しては Assessment Systems Corporation の Nathan Thompson 氏，CAT の心理学的側面については Northwestern University の Richard Gershon 教授と Northwest Evaluation Association の Steven Wise 氏に，それぞれ貴重な助言をいただいたことに厚く感謝の意を表します．またプログラム開発においては，Moodle UCAT のプログラムの開発に際し（株）Version2 様に，LRT-CAT のシミュレーションならびにテスト実施システムの開発・データ収集に際し（株）e ラーニング様に，ご尽力いただいたことに心より感謝いたします．研究で使用したテスト項目については，英検の過去問題のテスト項目の使用を許可してくださった日本英語検定協会に深く感謝の意を表します．事前テストに協力しくだった教育機関の方々と，何度もテストを受験してくれた多くの学生の皆さんに感謝いたします．

本研究の一部は科学研究費補助金基盤研究(C)(課題番号：22520590）を利用して行われました．また，早稲田大学 2010・2011・2012 年度国際会議論文発表補助費並びに 125 周年奨学金を利用することで，より多くの国際会議へ出席する機会を得られたことに感謝いたします，

最後に，この数年の間，研究活動に没頭することを暖かく見守ってくれた家族に感謝します．

参考文献

秋山實・木村哲夫・荘島宏二郎 (2011). LRTモデルに基づくCATの開発とシミュレーションによる特性解析. 日本テスト学会第9回大会発表論文抄録集, 146-147.

Albeck, H. (2012). Large scale adaptive testing of students grade (1-9). *International Association for Computer Adaptive Testing Conference 2012 Australia Conference Program*, 17.

Alderson, J.C. (2005). *Diagnosing foreign language proficiency: the interface between learning and assessment.* New York: Continuum.

Andrich, D. (1995). Review of the book Computerized Adaptive Testing: A Primer. *Psychometrika*, **4**, 615-620.

Andrich, D., Lyne, A., Sheridan, B., & Luo, G. (2010). RUMM2030. [Computer software]. Perth: RUMM Laboratory.

Assessment Systems Corporation & 4ROI. (2010). FastTEST Web. [Computer software]. http://www.fasttestweb.com/, St. Paul, MN.

Bergstrom, B. A., Lunz, M. E., & Gershon, R. C. (1992). Altering the difficulty in computer adaptive testing. *Applied Measurement in Education*, **5**, 137-149.

Binet, A., & Simon, Th. A. (1905). Méthode nouvelle pour le diagnostic du niveau intellectuel des anormaux. L'Année Psychologique, **11**, 191-244.

Birnbaum, A. (1968). Some Latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

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

Chang, H.-H., Qian, J., & Ying, Z. (2001). Alpha-stratified multistage computerized adaptive testing with beta blocking. *Applied Psychological Measurement*, **25**, 333–341.

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, **20**, 213–229.

Chang, H.-H., & Ying, Z. (1999). Alpha-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, **23**, 211–222.

張一平 (2007) 確信度テスト法と項目反応理論, 東京大学出版会, 東京.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge, England: Cambridge University Press.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, **16**(3), 297–334.

大学英語教育学会(JACET)学習ストラテジー研究会編 (2005). 言語学習と学習ストラテジー―自律学習に向けた応用言語学からのアプローチ, リーベル出版.

De Ayala, R. J. & Koch, W. R. (1986). A Computerized Implementation of a Flexilevel Test and Its

Comparison with a Bayesian Computerized Adaptive Test. (ERIC, ED269437).

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement,* **II**, 59-79.

Dunlea, J. (2009). The EIKEN can-do list: improving feedback for an English proficiency test in Japan. In L. Taylor & C.J. Weir (Eds.), *Studies in language testing* **31***: Language testing matters,* 245-262.

Dunlea, J. & Figueras, N. (2010). Replicating results from a CEFR test comparison project across continents. 7th Annual EALTA Conference.

Egberink, I. J. L., Meijer, R. R., Veldkamp, B. P., Schakel, L., & Smid, N. G. (2010). Detection of aberrant item score patterns in computerized adaptive testing: An empirical example using the CUSUM. *Personality and Individual Differences* **48**(8), 921–925.

Gershon, R. C. (1992). Test anxiety and item order: New concerns for item response theory. In M. Wilson (Ed.), *Objective measurement: Theory to practice.* Norwood, NJ: Ablex.

Green, B. F., Jr. (1983). The promise of tailored test. In: Wainer, H. and Messick, S. (Eds.), *Principles of modern psychological measurement* (pp. 69–80). Hillsdale, NJ: Laurence Erlbaum Associates.

Halkitis, P. N. (1993). A computer-adaptive testing algorithm. *Rasch Measurement Transactions* **6**:4, 245.

Hambleton, R. K. & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications.* Boston, MA: Kluwer-Nijhoff Publishing.

Han, K. T. (2009). A gradual maximum information ratio approach to item selection in computerized adaptive testing. *Research Reports 09–07*, McLean, VA: Graduate Management Admission Council.

Han, K. T. (2010). SimulCAT: Simulation software for computerized adaptive testing [computer program]. Retrieved March 20, 2010, from http://www.hantest.net/

Han, K. T. (2012). SimulCAT: Windows Software for Simulating Computerized Adaptive Test Administration. *Applied Psychological Measuremen*t, **36**(1), 64-66.

橋本貴充・植野真臣(2009). 局所従属性の指標に母数推定値が与える影響. 日本テスト学会第 7 回発表論文集, 80-83.

Hashimoto, T & Shojima, K. (2007). neutet. [Computer Software]. http://www.rd.dnc.ac.jp/~hashimot/neutet/

Häusler, J., & Sommer, M. (2008). The effect of success probability on test economy and self-confidence in computerized adaptive tests. *Psychology Science Quarterly,* **50**, 75-87.

Hinkelman, D., & Grose, T. (2004). Placement testing and audio quiz-making with open source software. Proceedings of CLaSIC 2004, 972-981.

Hornke, L. F. (1995). Item times in computerized testing – A new differential information. *European Journal of Psychological Assessment,* **11**, 108-109.

Hornke, L. F. (2000). Item response items in computerized-adaptive tests. *Psicologica*, **21**, 175-189.

Kimura, T. (2009). Construction of a Moodle-based placement test and possibility of a Moodle-based

computer adaptive test. *ARELE* **20**, 161-169.
木村哲夫 (2008a). Moodle を使ったテスティングとそのデータ分析, 金谷憲教授還暦記念論文集刊行委員会(編), 英語教育・英語学習研究:現場型リサーチと実践へのアプローチ:金谷憲教授還暦記念論文集, 247-258, 桐原書店.
木村哲夫 (2008b). Moodle を利用したテスト項目分析とアダプティブ・テスト開発の可能性, 第 34 回全国英語教育学会東京研究大会予稿集, 340-341.
木村哲夫 (2008c). 習熟度別クラス編成のための英語基礎力判定標準化テスト作成の試み, 第 12 回日本言語テスト学会プログラム, 21.
木村哲夫 (2009a). ニューラルテスト理論による英語プレイスメントテストの作成と評価. 関東甲信越英語教育学会研究紀要, **23**, 23-34.
木村哲夫 (2009b). NTTの実践的利用:2段階モデルによる英語プレイスメントテストの分析, 企画セッション「ニューラルテスト理論, 第7回日本テスト学会予稿集, 66-67.
木村哲夫 (2009c). 言語テストにおける段階評価の実際:入試とプレイスメントテストのデータ処理, 第13回日本言語テスト学会プログラム, 23.
木村哲夫 (2009d). Moodleによる英語プレイスメントテストのためのアイテムバンク構築, 第25回日本教育工学会全国大会予稿集, 581-582.
木村哲夫 (2010a) . 教員・学校間で共有する英語基礎力測定のアイテムバンク, 第36回全国英語教育学会 大阪研究大会予稿集, 140-141.
木村哲夫 (2010b). 段階評価のための項目分析:ニューラルテスト理論による分析, 大学英語教育学会第49回全国大会要綱, 170-171.
木村哲夫 (2011). 潜在ランク理論による診断的テスト結果の提示, 日本言語テスト学会第15回全国研究大会発表要綱, 31.
木村哲夫 (2012). 能力記述文による自己評価, 日本言語テスト学会第16回全国研究大会発表要綱, 37.
木村哲夫・永岡慶三. (2010a). ニューラルテスト理論による大問形式の英語読解問題のデータ分析, 第8回日本テスト学会予稿集, 102-105.
木村哲夫・永岡慶三 (2010b). Moodleによる小規模CAT構築に向けて1:アイテムバンクの拡充. JSET26講演論文集, 343-344.
木村哲夫・永岡慶三 (2011a). 潜在ランク理論に基づくコンピュータアダプティブテスト. 日本テスト学会第9回大会発表論文抄録集, 138-141.
木村哲夫・永岡慶三 (2011b). Moodleによる小規模CAT構築に向けて2:アイテムバンクの統合. JSET27講演論文集, 731-732.
木村哲夫・永岡慶三 (2012a). 潜在ランク理論に基づくコンピュータアダプティブテストーアルゴリズムの提案と検証ー. 日本テスト学会誌, **8**, 69-84.
木村哲夫・永岡慶三 (2012b). LRT-CATのアイテムバンク構築において望ましくない項目を除去する指針の提案. 日本テスト学会第10回大会発表論文抄録集, 180-183.

木村哲夫・永岡慶三 (2012c). Moodleによる小規模CAT構築に向けて3:アイテムバンクの検証. JSET28講演論文集, 193-194.

Kimura, T., Han, K., Kosinski, M., & Shoijma, K. (2012). A framework and approaches to develop an in-house CAT with freeware and open source software. *International Association for Computer Adaptive Testing 2012 Australia Conference Program*, 20.

Kimura, T. & Nagaoka, K. (2010). Towards the construction of item banks for moodle-based in-house computer adaptive English tests. *Pacific Rim Objective Measurement Symposium 2010 Malaysia Conference Program*, 47-48.

Kimura, T. & Nagaoka, K. (2011a). Reliability of Can-Do statements about EFL learners. *Pacific Rim Objective Measurement Symposium 2011 Singapore Conference Program*, 47-48.

Kimura, T. & Nagaoka, K. (2011b). Psychological aspects of CAT: How test-takers feel about CAT. *International Association for Computer Adaptive Testing Conference 2011 USA Conference Program*, 47.

Kimura, T. & Nagaoka, K. (2012a). Can difficulty of items be guessed intelligently without degrading CAT results? *Pacific Rim Objective Measurement Symposium 2012 China Conference Program*, 63.

Kimura, T. & Nagaoka, K. (2012b). Psychological aspects of CAT: seeking item selection rules which do not decrease test takers' learning self-efficacy and motivation. *International Association for Computer Adaptive Testing Conference 2012 Australia Conference Program*, 12.

Kimura, T. & Ohnishi, A. (2011). Moodle UCAT beta version: a computer-adaptive test module based on Rasch model. JALT CALL 2011.

Kimura, T. Ohnishi, A., & Nagaoka, K. (2012). Moodle UCAT: a computer-adaptive test module for Moodle based on the Rasch model, *The 5th International Conference on Probabilistic Models for Measurement Program*, 83.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, **2**(4), 359–375.

小泉利恵・飯村英樹 (2010). ニューラルテスト理論の特徴:古典的テスト理論・ラッシュモデリングとの比較から, 日本言語テスト学会研究紀要, **13**, 91-109.

Koyama, Y & Akiyama, M. (2009) Developing A Computer Adaptive ESP Placement Test Using Moodle. eLEARN2009　940-945Linacre, J.M. (1987). UCAT: a BASIC computer-adaptive testing program. MESA Psychometric Laboratory. (ERIC ED 280 895).

小山由紀江・木村哲夫 (2011). Neural Test Theory を使ったCan-do Statements の分析. 統計数理研究所共同研究リポート254, 科学技術コーパスの特徴語句抽出とその応用, 59-77.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, **2**(3), 151–160.

熊谷龍一 (2007). ニューラルテスト理論を離散変数型IRTとみなしたとき項目特徴を示す指標について. 第1回ニューラルテスト理論ワークショップ.

熊谷龍一 (2009). 初学者向けの項目反応理論分析プログラムEasyEstimationシリーズの開発. 日本テスト学会誌, **5**, 107-118.

Lange, R. (2007). Binary items and beyond: A simulation of computer adaptive testing using the Rasch partial credit model. In E. V. Smith & R.M. Smith (Eds.), *Rasch Measurement: Advanced and Specialized Applications* (pp.148-180). Maple Grove, MN: JAM Press.

Linacre, J.M. (1987). UCAT: a BASIC computer-adaptive testing program. MESA Psychometric Laboratory. (ERIC ED 280 895).

Linacre, J.M. (2000). Computer-adaptive testing: A methodology whose time has come. MESA Memorandum No 9.

Linacre, J.M. (2003). Size vs. Significance: Standardized Chi-Square Fit Statistic. *Rasch Measurement Transaction* **17**:1, p.918.

Linacre, J.M. (2006). Computer Adaptive Tests, Standard Errors and Stopping Rules. *Rasch Measurement Transaction* **20**:2, 1062.

Linacre, J. M. (2009). WINSTEPS [Computer software] (Ver. 3.70.1.1). Retrieved February 16th, 2009, from http://www.winsteps.com/, originally developed by Wright, B. D., and Linacre, J. M. (1998). Chicago: MESA Press.

Linacre, J.M. (2010). When to stop removing items and persons in Rasch analysis? *Rasch Measurement Transaction* **23**:4, 1241.

Lord, F. M. (1971). The self-scoring flexilevel test. Journal of Educational Measurement, **8**, 147-151.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M., & Novick. M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.

Lunz, M. E., & Bergstrom, B. A. (1994). An empirical study of computerized adaptive testing conditions. *Journal of Educational Measurement*, **31**, 251-263.

Magis, D. & Raîche, G. (2012). Random Generation of Response Patterns under Computerized Adaptive Testing with the R Package catR. *Journal of Statistical Software*, **48**(8), 1-31.

松宮功・荘島宏二郎 (2008). 下位テスト別潜在ランクを用いた中学生国語・数学の学力変化に関する考察. 日本テスト学会第6回大会発表論文抄録集, 132-133.

松宮 功・荘島宏二郎 (2009). ニューラルテスト理論を利用して作成する教科テストのCan-do table. 日本テスト学会第7回大会発表論文抄録集, 232-233.

Meijer, R. R. & Sijtsma, K. (1995). Detection of Aberrant Item Score Patterns: A Review of Recent Developments. *Applied Measurement in Education*, **8**(3), 261-272.

Mills, C. N. (1999). Development and introduction of a computer adaptive Graduate Record Examination General Test. In F. Drasgow & J .B. Olson-Buchanan (Eds.). *Innovations in computerized assessment* (pp. 117-135). Mahwah NJ:Erlbaum.

村木英治 (2011). 項目反応理論, 朝倉書店.

成田秀夫・荘島宏二郎・宇佐美 慧 (2010). ニューラルテスト理論を用いた大学生のジェネリックスキルを測定する試み. 第8回日本テスト学会抄録集, 60-61.

根岸雅史 (2011). CEFR-J 開発の経緯. *ARCLE REVIEW*, **5**, 38-52.

Nydick, S. W., & Weiss, D. J. (2009). A hybrid simulation procedure for the development of CATs. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.*

日本英語検定協会 (2007). 英検 Can doリスト. http://www.eiken.or.jp/about/cando/cando_02_0.html (accessed 2009.11.07).

野上康子・林 規夫 (2011). CASEC Can-Do リストの開発. 日本言語テスト学会第15回全国研究大会発表要綱, 37.

大友賢二 (1996). 項目応答理論入門, 大修館書店.

Ohtomo, K., Nakamura, Y., & Akiyama, M. (2002). Test Data Analysis Program (TDAP) Ver. 2.0 [Windows, Computer software]. In K. Ohtomo (ed.) & Y. Nakamaura, Test de Gengo Nouryoku ha Hakarerunoka: Gengo Test Data Bunseki Nyumon [Can Test Assess Language Ability? Introduction to Data Analysis of Language Test]. Tokyo: Kiriharashoten.

Ponsoda, V., Olea, J., Rodriguez, M.S., & Revuelta, J. (1999). The effect of test difficulty manipulation in computerized-adaptive testing and self-adapted testing. *Applied Measurement in Education*, 12, 167-184.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B. D. Wright. Chicago: The University of Chicago Press.

Reise, S. P., & Due, A. M. (1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement*, **15**(3), 217-226.

Rizopoulos, D. (2006). ltm: An R Package for Latent Variable Modeling and Item Response Theory Analyses. *Journal of Statistical Software*, **17**(5), 1-25.

斉田智里（2008）. ヨーロッパ言語共通参照枠(CEFR)による日本人大学生英語力診断の試み―英語教育達成目標への CEFR 適用可能性の一検討, *Jacet Journal*, **47**, 127-140.

佐野真 (2009). 相互情報量を用いた項目識別力の課題推定の検出. 日本テスト学会誌, **5**, 3-21.

芝祐順（1979）. 因子分析法, 第2版, 東京大学出版会.

芝祐順（編）(1991). 項目反応理論－基礎と応用－, 東京大学出版会.

芝祐順・渡部洋一・石塚智一（編）(1974). 統計用語辞典, 新曜社.

靜哲人 (2007). 基礎から深く理解するラッシュモデリング, 関西大学出版.

Shojima, K. (2007a). Neural test theory. *The International Meeting of the Psychometric Society 2007*, Tokyo, 160.

Shojima, K. (2007b). The graded neural test model: A neural test model for ordered polytomous data. *DNC Research Note*, RN07-03.

Shojima, K. (2008). Neural test theory: A latent rank theory for analyzing test data. *DNC Research Note*, 08-01.

Shojima, K. (2010). Exametrika [Computer software] (Ver. 4.3). Retrieved July 12, 2010 from http://www.rd.dnc.ac.jp/~shojima/exmk/

Shojima, K. (2012). Asymmetric triangulation scaling: Asymmetric multidimensional scaling for inter-item dependency structure. *Behaviormetrika*, **39**, 27-48.

荘島宏二郎 (2008a). ニューラルテスト理論－資格試験のためのテスト理論－, 平成 20 年度全国大学入学者選抜研究連絡協議会研究発表予稿集, 163-168.

荘島宏二郎 (2008b). The structural neurofield mapping: A latent rank model for multivariate data, 第 36 回日本行動計量学会大会発表論文抄録集, 179-180.

荘島宏二郎 (2009). 項目反応理論. 植野真臣・永岡慶三(編), eテスティング, pp.23-48, 培風館.

荘島宏二郎 (2010a). 古典的テスト理論―科学的対象としてのテスト得点―. 植野真臣・荘島宏二郎, 学習評価の新潮流, pp.37-55, 朝倉書店.

荘島宏二郎 (2010b). ニューラルテスト理論―学力を段階評価するための潜在ランク理論―, 植野真臣・荘島宏二郎, 学習評価の新潮流, pp.83-111, 朝倉書店.

STEP (The Society for Testing English Proficiency) (2007). The EIKEN Can-do List. http://stepeiken.org/download/sites/default/files/Eiken_CandoList_translation.pdf (accessed 2009.11.07).

Stocking, M. L., & Lewis, C. (1995). A new method of controlling item exposure in computerized adaptive testing. *Research Report* **95–25**. Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, **23(1)**,* 57–75.

Sympson, J. B. & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. In Proceedings of the 27th annual meeting of the Military Testing Association, (pp. 973–977), San Diego, CA: Navy Personnel Research and Development Centre.

竹内理 (2007). 自ら学ぶ姿勢を身につけるには―自主学習の必要性とその方法を探る―. TEACHING ENGLISH NOW VOL.8 SPRING, 2-5, 三省堂.

投野由紀夫 (2010). CEFR準拠の日本版英語到達指標の策定へ. 英語教育2010年10月増刊号, 60-63, 大修館書店.

Thissen, D., Chen, W-H., & Bock, R. D. (2003). MULTILOG 7 for Windows: Multiple-category item analysis and test scoring using item response theory [Computer software]. Skokie, IL: Scientific Software International, Inc.

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer, and et. al. (Eds.), *Computerized adaptive testing: a primer*（2nd ed.), 101-133. London: Lawrence Erlbaum Associates.

Thompson, N.A., & Weiss, D.J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research, and Evaluation,* **16**(1).

Tonidandel, S. & Quiñones, M. A. (2000). Psychological reactions to adaptive testing. *International Journal of Selection and Assessment*, **8**, 7-15.

Tonidandel, S., Quiñones, M. A., & Adams, A. A. (2002). Computer-adaptive testing: The impact of test characteristics on perceive3d performance and test takers' reactions. *Journal of Applied Psychology*, **87**, 320-332.

豊田秀樹 (1998). 共分散構造分析―構造方程式モデリング―[入門編], 書店.

豊田秀樹 (2002). 項目反応理論[入門編], 朝倉書店.

宇佐美慧 (2009). ニューラルテスト理論の応用可能性―方法論的課題の考察と多値型モデルの適用例―, 日本テスト学会誌, **5**, 66-79.

Van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, **63**, 201-216.

Van der Linden, W. J. (2010). Constrained Adaptive Testing with Shadow Tests. In van der Linden, W.J. & Glas, C.A.W. (Eds.) (2010). *Elements of adaptive testing* (pp.231-245). New York: Springer.

Veldkamp, B. P., & van der Linden, W. J. (2010). Designing item pools for adaptive testing. In van der Linden, W.J. & Glas, C.A.W. (Eds.) (2010). *Elements of adaptive testing* (pp.231-245). New York: Springer.

Veerkamp, W. J. J., & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing, *Journal of Educational and Behavioral Statistics*, **22**(2), 203–226.

von Davier, A. A. & Haberman, S. (2012). Comparability of Test Performance and Reported Scores in Multistage Testing. Keynote Speech at IACAT 2012 Conference, Sydney, Australia.

Weiss, D. J. (1973). The stratified adaptive computerized ability test (Research Report 73-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, **6**(4), 473–492.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology*, **53**(6), 774-789.

Weiss, D. J. & Guyer, R. (2010). *Manual for CATSim: Comprehensive simulation of computerized adaptive testing.* St. Paul MN: Assessment Systems Corporation.

Wright, B.D. (1988). Rasch model from Campbell Concatenation. *Rasch Measurement Transactions* **2**:1, 16.

Wright, B. D. and Douglas, G. (1975). Best test design and self-tailored testing. MESA Memorandum No. 19. Department of Education, Univ. of Chicago.

Yao, T. (1991). CAT with a poorly calibrated item bank. *Rasch Measurement Transactions* **5**:2, 141.

Zickar, M. J., Overton, R. C., Taylor, L. R., & Harms, H. J. (1999). The development of a computerized adaptive selection system for computer programmers in a financial services company. In F. Drasgow

and J. B. Olsen (Eds.), *Innovations in computerized assessment* (pp. 7-33). Mahwah, NJ: Erlbaum.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R.D. (2003). BILOG-MG 3 for Windows: Multiple-group IRT analysis and test maintenance for binary items [Computer software]. Skokie, IL: Scientific Software International, Inc..