適応型テストの現状－能力測定と診断

教育心理学コース　平　井　洋　子

A Review of the Studies on Adaptive Testing :
Ability Measurement and Diagnosis

Yoko HIRAI

Adaptive testing, in contrast to the conventional tests in which all the examinees take the same set of test items, presents an examinee only the items that are informative on the examinee's ability. Therefore, in adaptive testing, examinees don't have to answer items that are too easy or too difficult to them. In 70s to 80s, research concerns concentrated on its psychometric characteristics ; the precision and the efficiency of measurement. It was found that the adaptive testing, typically computerized and based on the item response theory, can measure ability in about half the test length without compromising the precision of measurement. In last ten years, however, research efforts were added to make the (computerized) adaptive testing more useful, more valid, examinieefriendly, and diagnostic. The efforts to elaborate the IRT model, to balance the content domain, to investigate the individual difference, and to incorporate the artificial intelligence or expert systems in this context are also reviewed.

目　　次

Ⅰ．序　論
Ⅱ．CATの構成要素
　A．項目反応モデル
　B．項目プール
　C．初めの項目の困難度
　D．項目選択ルール
　E．採点方法
　F．終了基準
Ⅲ．IRT－CATの実証的研究
Ⅳ．多様な展開
　A．項目反応モデルの多値化
　B．被験者の心理面への配慮
　C．テスト内容のきめ細かな指定
　　1．テストレット（testlet）
　　2．線形計画法による最適化
　D．人工知能からのアプローチ
　　1．自然言語処理
　　2．知的項目選択
Ⅴ．結語：診断的評価に向けて

Ⅰ．序　論

従来の大量実施を想定したテストでは，被験者にさまざまな能力レベルの人がいることが予想されたので，幅広い困難度のテスト項目が用意された。そのため，能力の高い被験者には易しすぎる項目が含まれたり，逆に能力の低い被験者には難しすぎる項目が含まれるといった問題があった。易し過ぎたり難しすぎる項目は，テストを実施しなくても被験者の成績（正答／誤答）が予測できるため，実施による情報の追加はほとんどない。このような無駄を省くために，被験者の能力レベルにあった項目だけを受験させようとするのが，適応型テストである。

適応型テストの基本的な考え方は，賢明な試験官が個別テストで行うことを自動的に遂行する，というものである（Wainer, 1990)。すなわち，被験者がある項目に正答すれば，次はより難しい項目を課し，誤答すれば次はより易しい項目を課して，効率よく能力レベルに関する情報を集めるわけである。

最もシンプルな適応型テストは，２段階テストと呼ばれるものであろう。これは，まず被験者全員に同じ１段階めのテストを受けさせ，その成績に応じて，いくつか

の2段階めのテストのうち適切な困難度のものを割り振る方式である。適応させる手続きは1段階めから2段階めに移るときの1回のみである。この適応の手続きが複数回になったものが多段階テストである。多段階テストには，その適応・分岐のさせかたによってピラミッド型(ツリー型)，多層型などがある。多段階テストにはさらに，フレキシレベル方式も含めることができよう。これらのテスト方式についてはLord (1980) に詳しく述べられている。

適応型テストの発達に大きな役割を果たしたのが，項目反応理論 (Item Response Theory, IRT) というテスト理論である (Birnbaum, 1968; Hambleton & Swaminathan, 1985; Lord, 1980)。それ以前のテスト理論がテストを単位としていたのに対し，項目反応理論は項目を単位とする。適応型テストのように被験者によって受ける項目が異なる場合でも，同じ尺度上で能力を比較することができる (Wainer, 1990)。項目反応理論の他の利点としては，能力と項目困難度が同じ尺度で計量されるため適切な困難度の項目が選びやすい点，測定精度をテストの終了基準にできる点，があげられる (Weiss, 1982)。測定精度が終了基準になるということは，測定精度をテスト実施者側がコントロールできることを意味する。

さて適応型テストでは，適応させるたびに被験者の回答を何らかの形で採点しなければならない。紙筆形式 (paper and pencil mode) のテストでは，被験者に自己採点させたり，次に回答すべき項目を指示するなどの工夫が必要になる (芝・野口・南風原，1978)。もともと項目反応理論による能力推定にはコンピュータによる計算が必須であったが，コンピュータの発達に伴い，テスト自体をコンピュータ化し，この採点・適応手続きをテストに組み込もうとするのは自然の流れであったろう (Computerized Adaptive Testing; CAT)。1980年代に入ると，項目反応理論に基づくCAT (IRT－CAT) が精力的に開発・実施された。

II. IRT－CATの構成要素

典型的なCATでは，被験者が項目に回答するごとにその能力を推定し直し，新しい能力推定値に最も適切な項目を未実施の項目群から探して実施する。被験者が正答すれば新しい能力推定値は上り，次にはより難しい項目が提示される。誤答すれば新しい能力測定値は下がり，より易しい項目が提示されることになる。

CATの手続きは以下の構成要素から成り立っている (Weiss & Kingsbury, 1984)。

a) 項目反応モデル　b) 項目プール　c) 初めの項目の困難度　d) 項目選択ルール　e) 採点方法　f) 終了基準。それぞれいくつかの方式があり，テストの目的に応じて個別に仕様を決めていくことになる。

A. 項目反応モデル

最も多く用いられたモデルは，3パラメタ・ロジスティックモデル (3PL) であろう (Green, Bock, Humphreys, Linn, & Reckase, 1984)。このモデルは，あて推量による正答確率を考慮した，多枝選択式項目に適したモデルである。この他，あて推量を考慮しない2パラメタ・ロジスティックモデル (2PL)，さらに項目の識別力を一定と仮定したRaschモデルなどもよく用いられている。これらの項目反応モデルの詳細はHambleton & Swaminathan (1985) を参照されたい。

B. 項目プール

適応型テストでは，被験者の能力レベルに合った項目を次々と提示できなければならない。さまざまな能力レベルの被験者に対応するには，多くの項目をプールしておく必要がある。Urry (1977) は，3PLモデルの場合，項目プールの条件として　a)識別力パラメタの値がいずれも0.8以上　b)困難度パラメタの値が広く一様に分布 ($-2.0<\theta<2.0$)　c)あて推量パラメタの値がいずれも0.3以下　d)項目プールサイズは最低100項目　をあげている。また，Jensema (1977) は，シミュレーションによって，項目プールに必要とされる条件について調べた。その結果，高い識別力，低いあて推量の項目が十分多くあること，困難度が一様に分布していることが大切で，それが満たされていれば項目プールのサイズは絶対的な条件にはならないと結論づけた。とはいえ，実際の運用にあたっては75項目以上が望ましいと指摘している。

一般に適応型テストでは，従来のテストに比べて1人の被験者が受験する項目数が少ない。従って，欠点のある項目が及ぼす影響が相対的に大きくなる。テストの妥当性・公平性を保つために，事前に入念な項目分析が必要とされる (Wainer, 1990)。Ree (1981) は項目パラメタの推定精度に着目し，シミュレーションを行った結果，3パラメタ・ロジスティックモデルで200項目のパラメタを推定するなら，2000人のデータが必要だと指摘している。

C．初めの項目の困難度

原理的には，どのレベルの困難度から始めてもよい。他に何も情報がなければ，被験者母集団の平均レベルの項目を第1問目にするのが，得られる情報が最も多いという点で最適である。一方，被験者の学年や履修科目など，能力に関わる情報が得られている時は，それを考慮した項目提示が考えられよう（Wainer, 1990）。テストバッテリー中の他のサブテストの結果が得られていれば，(重) 回帰によって，初期レベルを推定することも考えられる（Weiss, 1982）。

被験者の能力推定値は，適応型テストが進行するにつれ被験者の能力ベルに収束する。従って，初めの項目の困難度が真の能力レベルと掛け離れていても，結果に重大な影響を及ぼすことはない（Wiess & Kingsbury, 1984）。初めの項目が被験者の能力レベルと一致していたときのメリットは，テストが数項目だけ短くて済むことである（Weiss, 1982）。

D．項目選択ルール

ある項目に対する回答が得られた後，未実施の項目から次にどれを選んで提示するかという方法が項目選択ルールであるが，これには大きく分けて2通りの方法がある。最大項目情報量方式と，ベイズ流最小期待事後分散方式である。

最大項目情報量方式とは，現在の能力推定値に対する項目情報量が最大の項目を選ぶ方法である。この方式では，現在の能力推定値に近い項目困難度を持ち，しかも項目識別力が大きい項目が選ばれる傾向がある。最大項目情報量方式の特徴は，各被験者について，能力推定の標準誤差が最小になることである（Weiss, 1982）。反面，良い能力推定値が得られない時のための工夫も必要となる。例えば柴山・野口・芝・鎌原（1987）は，初めの6項目では正答の後は項目困難度にして＋1だけ難しい項目を提示し，誤答の後は－1だけ易しい項目を提示する固定ステップサイズをとることで，能力推定の非収束・不安定さを回避している。またDodd（1990）は，固定ステップサイズより可変ステップサイズの方が非収束ケースが少なかったとして，可変ステップサイズを奨めている。

ベイズ流最小期待事後分散方式（Jensema, 1977; Owen, 1975）は，現在までの被験者の反応を事前情報とし，次の項目に正答したときの事後分散と誤答した時の事後分散の期待値を，未実施の項目すべてについて計算する。その事後期待分散が最小の項目を次に提示する方式である。Owenの方法は，反復計算が不要なので1980年代にはよく用いられたが，近似的な方法でもあり，最近はあまり用いられていない（Wainer, 1990）。

E．採点方法

CATでは，一般に被験者の回答が得られるごとに能力を推定し直す。この方法にも大きく2通りがある。最尤推定法とベイズ推定法である。これら推定法の技術的な面についてはBaker（1992）を参照されたい。

最尤推定法の問題は，被験者が正答し続けたり誤答し続けた時に，能力推定値が有限にならないことである。従って，1番目の項目から2番目の項目に移る時はもちろん，初めの数項目は有限の能力推定値が得られない可能性がある。一方，ベイズ推定法の問題は，事後分布のモード・期待値がバイアスを持つ点である（McBride, 1977; Weiss & McBride, 1984）。これらの推定量は一致性を持つ（Bock & Mislevy, 1982; Owen, 1975）ので，無限にテストを実施すれば真の値に一致するが，現実にはそのようなことはありえない。このバイアスの正体は，事後推定値の事前平均への回帰である。逆に，この回帰効果によって，事後推定値が無限大に発散せず，安定することになる。

これらの問題に対し，いくつかの提案がなされている。たとえば，初めから最尤推定法で行う場合はステップサイズを導入するなどして，適切な推定値が得られるまで何らかの暫定的推定値を用いる（前節を参照）。初めの数項目は項目選択のためだけベイズ推定を行い，最尤推定ができるようになったら切り替える（Weiss, 1982）。事前分布を矩形分布とする（McBride, 1977）などである。

しかしながら最終的な能力推定では，実施項目数が増えるにつれて事前分布の貢献が相対的に小さくなるので，結果的に推定方法の違いはあまり問題にならない（Weiss, 1982）。例えば20項目も実施すると，両者はほぼ同じ推定値をもたらす（Wainer, 1990）。

F．終了基準

CATをいつ終了するかについても，いくつかの方法がある。Wainer（1990）は，

① 目標の測定精度が得られたとき
② 決められた項目数に達したとき
③ 決められた時間に達したとき

の3方法をあげている。このうち，①の方法は，全員が同じ測定精度になるという点で好ましい。②の方法は最も実施が簡単だが，能力の非常に高い被験者，非常に低い被験者の測定精度が低くなる恐れがある。また，③の

方法は主としてスピードテスト用の基準である。Samejima (1977) はまた，能力推定値の変化がある基準以内になったときを，終了基準にあげている。

これらの終了基準は，単独または複合で用いられる。現実的には，テストに何らかの精度を求めることが多いので①を，そして項目プールを使い切ったときのために②を同時に設定しておくのが一般的である。

CATに関する総合的な議論はWainer (1990)，Weiss (1982) が詳しい。また，Green, Bock, Humphreys Linn, & Reckase (1984) には，CATシステムを評価するための指針が社会的な問題も考慮しながら述べられている。

III. IRT－CATの実証的研究

IRT－CATの実証的研究は数多くなされている。その一部まとめたものが表1である。これらの研究により，CATでは高い信頼性が得られること (服部, 1989, 1990)，しかも従来の項目固定型テストの約半分の項目数で同じだけの測定精度が得られること (Lord, 1977; Urry, 1977)，外的基準との相関（妥当性）も下がらないこと (McBride & Martin, 1983)，などが示された。

また，習得テストに関しては，尤度比検定を継次的に行う方法 (Reckase, 1983) や信頼区間を用いて幅広い能力層を同じ精度で測定する方法 (Kingsbury & Weiss, 1983) の有効性が示された。従来のテスト理論では，習得テストはカッティング・ポイントに困難度を合わせたテスト (peaked test) がよいとされていたが，IRT－CATでは被験者に最適の項目を提示することで，より効率・精度の良い測定ができるようになった。

Schoonman (1989) は，オランダ鉄道会社 (Dutch Railways) の採用試験用に，GATB (General Aptitude Test Battery) をCAT化する長期プロジェクトを記録した特筆すべき文献である。項目反応モデルの選択，適応アルゴリズムの検討，項目プールの作成，シミュレーション実験，実際に運用しての検討など，CATの具体的な作成過程が詳細に記録されている。また，被験者の人格的側面や反応時間データの利用可能性の他，経済的側面・ハードの側面などが，実際的な視点から議論されている。

IV. 多様な展開

このようにIRT－CATは，能力測定を効率良く，精度良く遂行できることが示された。しかし，1980年代半ばごろから，それまでのCATでは正答／誤答データのみしか扱えない，被験者の心理的側面を無視している，項目選択時に項目の内容を指定できない，などの問題が指摘されるようになった。これらの問題を克服するべく，多様な研究・提案がなされるようになった。

A. 項目反応モデルの多値化

現実のテストデータは，部分点を含む多値採点であることが多い。項目反応理論で多値データを扱うモデルには，段階反応モデル (Samejima, 1969)，部分反応モデル (Masters, 1982)，評定尺度モデル (Andrich, 1978)，名義反応モデル (Bock, 1972) などがある。これらのモデルのCAT化に関しては，De Ayala, Dodd, Kochらが精力的に研究を行っている（表2）。彼らの一連の研究の結果わかったことは，多値モデルのCATでは一般に項目プールのサイズが30項目程度ですむこと，3PLモデル（2値）とほとんど変らない能力推定値が得られること，それを3PLモデルより少ない項目数で実現できること，などである。これらの成果は，1項目の情報量が広い範囲にわたって多いという，多値モデル一般の特徴によってもたらされたものである (Dodd, 1990)。なおDodd, Koch, & De Ayala (1989) には，段階反応モデルのCATを構築するための指針がまとめられている。

B. 被験者心理面への配慮

従来行われていた測定の精度や効率の追求は，主として試験者の側に立った研究であった。しかし最近では，被験者の心理的負担を考慮した研究もなされている。

CATが被験者にどのように受け止められているかについて，アンケートを取った研究がいくつかある（藤森, 1995；服部, 1990；Schonman, 1989；柴山・野口・芝・鎌原, 1987)。その結果を見ると，テストのやりやすさ，画面の見やすさ，回答入力のしかたなど実施方法についての抵抗感はないようである。一方，項目の難しさについては，測定論的に最適な項目（正答確率50%）を難しいと感じる被験者が多かった。例えば藤森 (1995) は，被験者がちょうどよい困難度と感じた場合の正答率は80%程度であることを報告している。

被験者の心理的負担，特にテスト不安を軽減するため，Rocklin & O'Donnell (1987) は，自己調整テスト (self-adapted testing) を提案した。これは，困難度に応じて項目を層化しておき，被験者がある項目に回答するたびに正誤を知らせ，被験者に次の項目の困難度を選ばせる，という方式である。この方式で実施したところ，被験者はテストが進むにつれて難しい項目を選ぶように

表1　IRT－CATの実証的研究

	テスト内容		モデル	項目プール	項目選択	能力推定	終了基準	注
服部（1989, 1990）	能力テスト	語彙理解力	Rasch	346項目	最大情報量	最尤推定	標準誤差0.4以下または30項目	服部（1990）は実際にCATを運用
Lord（1977）		シミュレーション	2PL	182項目	最大情報量	最尤推定	25項目	Broad-Range Tailored Test
McBride & Martin（1983）		言語能力	3PL	150項目	Owen（1975）	Owen（1975）	30項目	実際にCATを運用
Schoonman（1989）		GATB	Rasch	60～123項目	Owen（1975）	事後期待値	事後分散0.5以下または制限時間	4つのサブテストについてCATを作成，うち2つを実際に運用
柴山・野口・芝・鎌原（1987）		語彙理解力	2PL	241項目	最大情報量	最尤推定	標準誤差0.2以下または20～60項目	実際にCATを運用
Urry（1977）		言語能力	正規累積（3パラメタ）	200項目	Owen（1975）	Owen（1975）	標準誤差0.26～0.55まで8段階	信頼性と妥当性を検討
Waters（1977）		言語アナロジー	正規累積（2パラメタ）	244項目	多層適応型	不明	5項目以上回答ししかも25%以下の正答率の層の出現	多層適応アルゴリズムによるCATを作成
Reckase（1983）	習得テスト	シミュレーション	Rasch 3PL	72項目	最大情報量	最尤推定	尤度比検定または20項目	Wald の Sequential Probability Testによる
Kingsbury & Weiss（1983）		シミュレーション	3PL	100項目	最大情報量	Owen（1975）	ベイズ信頼区間にカッティングポイントが含まれなくなるまで	カッティングポイントの近くの能力の被験者は点推定値を用いる

表2　De Ayala, Dodd, Kochらによる多値モデルCATの研究

	IRTモデル	項目プール	項目選択	能力推定	終了基準	注
De Ayala (1989)	名義反応モデル	50項目	最大情報量	最尤推定	標準誤差が0.45以下または30項目	3PLと比較 実データにもとづくシミュレーション
De Ayala (1992)	名義反応モデル	90項目	最大情報量	事後期待値	30項目固定	3PLと比較 完全シミュレーション
De Ayala, Dodd, & Koch (1988)	段階反応モデル 名義反応モデル	50項目	最大情報量	最尤推定	標準誤差が0.44以下または30項目	名義反応モデルでは終了基準が0.45以下 実データにもとづくシミュレーション
De Ayala, Dodd, & Koch (1990)	部分反応モデル 段階反応モデル	55項目	最大情報量	最尤推定	標準誤差が0.10, 0.25, 0.30以下，または20項目	完全シミュレーション
Dodd (1990)	評定尺度モデル	24〜39項目	最大情報量 最近項目	最尤推定	残り項目の情報量が0.44以下，または標準誤差が0.30以下，または20項目	実データにもとづくシミュレーション および完全シミュレーション 0.44は$-3<\theta<3$の平均項目情報量
Dodd, Koch, & De Ayala (1993)	評定尺度モデル	30, 60項目	最大情報量	最尤推定	残り項目の情報量が0.45以下，または標準誤差が0.30以下，または20項目	完全シミュレーション 項目プールの特性をいろいろに変化

なったという。

CATでは，通常項目を逆上ぼって回答訂正できないが，これも被験者の心理的負担につながる恐れがある。実際CATでは，同じ時間内に回答する項目数が紙筆テストに比べてかなり少ないという報告もある(Schoonman, 1989)。Lunz, Bergstrom, & Wright (1992) は，CAT終了後，被験者に回答を見直させる実験を行った。その結果，見直しによる能力推定値の上昇は無視できる程度しかなかったとして，人生を左右するようなテストでは見直すことを認めるべきだとしている。

C．テスト内容のきめ細かな指定

従来のCATでは，提示される項目は被験者の能力推定値と項目の統計的特徴だけから決まっていた。いま四則演算の項目がバランス良く入った項目プールがあり，除算の項目が平均的に高い項目識別力をもっていたとする。このとき従来のCATでは，除算の項目が優先的に提示され，意図した四則演算のテストとは異なる内容のテストになる。このような事態を避けるためには，項目選択アルゴリズムに提示項目の内容（下位領域）に関する何らかに制約を入れなければならない。

1．テストレット (testlet)

テストレットとは1つの内容領域に属する項目のセットで，テストレット内での項目の構造や項目数はあらかじめ決められている。また，テストレット内では項目は互いに従属していてもよいが，テストレット間では独立とする。この概念は，従来のCATで生じていた項目提示の文脈効果（順序効果，他の項目が情報源，内容のアンバランス）や提示順序のコントロールのために提案された (Wainer & Kiely, 1987)。CATを項目単位でなくテストレット単位で適応させることで，文脈効果や提示順序の問題がかなり軽減される。また，内容のコントロールも容易になる。

テストレットの心理測定的側面の研究にはSireci, Thissen, & Wainer (1991), Thissen, Steinberg, & Mooney (1989) があり，応用的研究にはLewis & Sheehan (1990), Sheehan & Lewis (1992) がある。

2．線形計画法による最適化

テストレットは概念的なもので，項目選択アルゴリズムへの制約の入れ方については具体的に言及されていない。制約の入れかたの技術的な解決方法としては，線形計画法による最適化が一般的であろう。線形計画法では幅広い制約が扱え，テストレットによる制約もその一部とすることができる。Adema (1990), Adema & van der Linden (1989), Boekkooi-Timminga (1990), Theunissen (1985, 1986) らは，項目選択への制約を多次元ナップサック問題としてとらえ，項目情報関数や内容領域，困難度，項目数などに制約が加わったときの解法を示した。彼らの解法は必ずしも適応型テストを念頭に置いたものではないが，Foong & Lam (1991) は2段階テストをCAT化し，多次元ナップサック法によって項目選択を行ってその有効性を確認している。また，Stocking & Swanson (1993) は重み付き偏差モデル (weighted deviation model) を応用して，解が見つからない時のために，制約を「要求」と見なして次善解を求める手続きを示した。

D．人工知能からのアプローチ

教育分野はあまり構造化されてなく，人工知能的手法の応用は遅れぎみである (McArthur, 1985; McArthur & Choppin, 1984) が，今後の発展が楽しみともいえる。

1．自然言語処理

従来のCATは，多枝選択式項目やLikert尺度などの項目形式を扱っていた。しかし，項目から得られる情報は記述式項目の方が多い。CASIPは，自由記述 (open-end) 項目を採点し，結果によって異なる質問を提示したり，回答の説明を求めたりできる，対話型のCAIシステムである (Anbar, 1986, 1991; Henning, Anbar, Helm, & D'Arcy, 1993)。得点の与え方や次の項目への移行は，比較的自由に試験者が設定することができる。このシステムの採点・評価部分はまさに自然言語を扱うCATといえるが，生み出されたのが医学教育の分野であり，心理測定の枠組みでなかったところが興味深い。

2．知的項目選択

被験者がある項目に誤答したとき，その項目が必要とするスキルを用いる上位の項目にも誤答する可能性が高い。逆にある項目に正答したとき，その下位スキルを測る項目には正答する可能性が高い。このようにテスト中の各項目が求める知識・スキル間の関係が構造化されるとき，それを利用して被験者の個々の知識・スキルの習得状況を探ることができる (Lesgold, Bonar, & Ivill, 1987; Spineti & Hambleton, 1977; 許・繁桝, 1990)。またLuk (1991) はエキスパートシステムによる習得／未習得の判定と，従来のCATによる判定とを比較し，CATよりも正確かつ効率のよい習得判定ができたと報告している。

V．結語：診断的評価に向けて

項目反応理論に基づくCATは測定効率が非常によく，

選抜や能力測定に大きな力を発揮する。1980年代は，それまでの大量実施の紙筆テストに比べ，いかに精度・効率のよい測定を実現するかが研究された時期だったといえる。だが，世の中のテストは必ずしも選抜や序列化を目的としたものだけではない。特に教育場面での活用を考えると，個々のスキルや知識の獲得状況を明らかにする診断的テストの果たす役割は大きい。質的データを広く扱う名義反応モデルは，選択枝が診断的情報を含む正解のない項目も扱うことができるので，診断的テストへの発展の可能性を秘めている。とはいえ，項目反応理論は基本的に1次元の尺度であり，習得状況に関する細かで質的な診断的情報は得にくい。

これに対し，前節の人工知能からのアプローチはCAIなどの文脈から生まれたものであり，本質的に適応的，診断的性格を持ったテストを構成する。測定論的に見ても，数理モデルや心理測定の知見を取り入れ，精緻な基盤を持つものが多い。さらに1990年代に入るごろからは，認知心理学の研究成果と心理測定理論とを結び合わせて診断的評価を目指そうとする動きも現れてきた(Frederiksen, Glaser, Lesgold, & Shafto, 1990)。Nichols, Chipman, & Brennan (1995) には，ベイズ統計学を用いた学習モデリング，概念ネットワーク，心理測定的モデリングに関する研究が幅広く収められている。

コンピュータは紙筆テスト時代の項目形式を扱えるのはもちろん，反応時間を記録したり，動画を提示することができる。ビデオ等を利用したシミュレーションテストや，記憶のテストを行うことも可能である。こうした新しい項目形式に工学・認知科学の知見を結合することで，CATはより豊かな評価システムに発展していくと思われる。

(指導教官　渡部洋教授)

引用文献

Adema,J.J. 1990 The construction of customized two-stage tests. *Journal of Educational Measurement, 27*, 241-253.

Adema,J.J. & van der Linden,W.J. 1989 Algorithms for computerized test construction using classical item parameters. *Journal of Educational Statistics, 14*, 279-290.

Anbar,M. 1986 *CASIP … a novel authoring tool for open ended natural language CAI.* Paper presented at the International Conference of the Association for the Development of Computer-Based Information Systems (Washington,DC. November 10-13, 1986.).

Anbar,M. 1991 Comparing assessments of students' knowledge by computerized open-ended and multiple-choice tests. *Academic Medicine, 66*, 420-422.

Andrich,D. 1978 A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

Baker,F.B. 1992 *Item response theory: Parameter estimation techniques.* New York: Marcel Dekker.

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

Bock,R.D. 1972 Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Bock,R.D. & Mislevy,R.J. 1982 Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431-444.

Boekkooi-Timminga,E. 1990 A cluster-based method for test construction. *Applied Psychological Measurement, 14*, 341-354.

De Ayala,R.J. 1989 A comparison of the nominal response model and the three-parameter logistic model in computerized adaptive testing. *Educational and Psychological Measurement, 49*, 789-805.

De Ayala,R.J. 1992 The nominal response model in computerized adaptive testing. *Applied Psychological Measurement, 16*, 327-343.

De Ayala,R.J., Dodd,B.G., & Koch,W.R. 1988 *A comparison of the nominal and graded response models in computerized testing* Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans LA, April 5-9, 1988).

De Ayala,R.J., Dodd,B.G., & Koch,W.R. 1990 *A comparison of the partial credit and graded response models in computerized adaptive testing* Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).

Dodd,B.G. 1990 The effect of item selection procedure and stepsize on computerized adaptive attitude measurement using the rating scale model. *Applied Psychological Measurement, 14*, 355-366.

Dodd,B.G., Koch,W.R., & De Ayala,R.J. 1989 Operational characteristics of adaptive testing procedures using the graded response model. *Applied Psychological Measurement, 13*, 129-143.

Dodd,B.G., Koch,W.R., & De Ayala,R.J. 1993 Computerized adaptive testing using the partial credit model: Effects of item pool characteristics and different stopping rules. *Educational and Psychological Measurement, 53*, 61-77.

Foong,Y.Y. & Lam, T.L. 1991 *The use of the graded response model in computerized adaptive testing of the attitudes to science scale* Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991).

Frederiksen,N., Glaser,R., Lesgold,A., & Shafto,M.G. 1990 *Diagnostic monitoring of skill and knowledge acquisition* Hillsdale,NJ: Lawrence Erlbaum.

藤森進 1995 テスト項目の心理的に最適な困難度水準の研究 心理学研究 第65巻 446-453.

Green,B.F., Bock,R.D., Humphreys,L.G., Linn,R.L., & Reckase,M.D. 1984 Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*, 347-360.

Hambleton,R.K. & Swaminathan,H. 1985 *Item response theory:*

Principles and applications Boston,MA: Kluwer-Nijhoff.
服部環 1989 調整テスト方式により中学生の語彙理解力を測定する試み 日本教育工学雑誌 第13巻 129-137.
服部環 1990 個人差に応じたテスト方式による語彙理解力の測定 教育心理学研究 第38巻 445-454.
Henning,G., Anbar,M., Helm, C.E., & D'Arcy,S.J. 1993 Computer-assisted testing of reading comprehension: Comparisons among multiple-choice and open-ended scoring methods. In D.Douglas & C.Chapelle (eds.), *A new decade of language testing research: Selected papers from the 1990 language testing research colloquium* Alexandria, VA: TESOL.
Jensema, C.J. 1977 Bayesian tailored testing and the influence of item bank characteristics. *Applied Psychological Measurement, 1*, 111-120.
Kingsbury,G.G. & Weiss,D.J. 1983 A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D.J.Weiss (ed.), *New horizons in testing: latent trait test theory and computerized adaptive testing* New York: Academic Press.
Lesgold,A., Bonar,J, & Ivill,J. 1987 *Toward intelligent systems for testing* Technical Report. Learning Research and Development Center, University of Pittsburgh, Pittsburgh, PA.
Lewis,C. & Sheehan,K. 1990 Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14*, 367-386.
Lord,F.M. 1977 A broad-range tailored test of verbal ability. *Applied Psychological measurement, 1*, 95-100.
Lord,F.M. 1980 *Applications of item response theory to practical testing problems* Hillsdale, NJ: Lawrence Erlbaum.
Luk,H. 1991 *An empirical comparison of an expert systems approach and an IRT approach to computer-based adaptive mastery testing* Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991).
Lunz,M.E., Bergstrom,B.A., & Wright,B.D. 1992 The effect of review on student ability and test efficiency for computerized adaptive tests. *Applied Psychological Measurement, 16*, 33-40.
Masters,G.N. 1982 A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.
McArthur,D.L. 1985 *Computerized diagnostic testing: Problems and possibilities* CSE Report No. 255. Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.
McArthur,D.L. & Choppin,B.H. 1984 Computerized diagnostic testing. *Journal of Educational Measurement, 21*, 391-397.
McBride,J.R. 1977 Some properties of a Bayesian ability testing strategy. *Applied Psychological Measurement, 1*, 121-140.
McBride,J.R. & Martin,J.T. 1983 Reliability and validity of adaptive ability tests in a military setting. In D.J.Weiss (ed.), *New horizons in testing: latent trait test theory and computerized adaptive testing* New York: Academic Press.
Nichols,P.D., Chipman,S.F., & Brennan,R.L. 1995 *Cognitively diagnostic assessment* Hillsdale, NJ: Lawrence Erlbaum.
Owen,R.J. 1975 A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70*, 351-356.
Reckase,M.D. 1983 A procedure for decision making using tailored testing. In D.J.Weiss (ed.) *New Horizons in Tesitng: Latent trait test theory and computerized adaptive testing* New York: Academic Press.
Ree,M.J. 1981 The effects of item calibration sample size and item pool size on adaptive testing. *Applied Psychological Measurement, 1*, 11-19.
Rocklin,T. & O'Donnell,A.M. 1987 Self-adapted testing: A performance-improving variant of computerized adaptive testing. *Journal of Educational Psychology, 79*, 315-319.
Samejima,F. 1969 *Estimation of latent ability using a response pattern of graded scores* Psychometrika Monograph, No.17.
Samejima,F. 1977 A use of the information function in tailored testing. *Applied Psychological Measurement, 1*, 233-247.
Schoonman,W. 1989 *An applied study on computerized adaptive testing* Amsterdam: Swets & Zeitlinger.
Sheehan,K. & Lewis,C. 1992 Computerized mastery testing with nonequivalent testlets. *Applied Psychological Measurement, 16*, 65-76.
芝祐順・野口裕之・南風原朝和 1978 語彙理解力測定のための多層適応形テスト 教育心理学研究 第26巻 229-238.
柴山直・野口裕之・芝祐順・鎌原雅彦 1987 最適化テスト方式による語彙理解力の測定 教育心理学研究 第35巻 363-367.
Sireci,S.G., Thissen,D., & Wainer,H. 1991 On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*, 237-247.
Spineti,J.P. & Hambleton,R.K. 1977 A computer simulation study of tailored testing strategies for objective-based instructional programs. *Educational and Psychological Measurement, 37*, 139-158.
Stocking,M.L. & Swanson,L. 1993 A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17*, 277-292.
Theunissen,T.J.J.M. 1985 Binary programming and test design. *Psychometrika, 50*, 411-420.
Theunissen,T.J.J.M. 1986 Some applications of optimization algorithms in test design and adaptive testing. *Applied Psychological Measurement, 10*, 381-389.
Thissen,D., Steinberg,L., & Mooney,J.A. 1989 Trace lines for testlets: A use of multiple-categorical-response-models. *Journal of Educational Measurement, 26*, 247-260.
Urry,V.W. 1977 Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement, 14*, 181-196.
Wainer,H. 1990 *Computerized adaptive testing: A primer* Hillsdale,NJ: Lawrence Erlbaum.
Wainer,H. & Kiely,G.L. 1987 Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-201.
Waters,B.K. 1977 An empirical investigation of the stratified adaptive computerized testing model. *Applied Psychological Measurement, 1*, 141-152.
Weiss,D.J. 1982 Improving Measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*, 473-492.
Weiss,D.J. & Kingsbury,G.G. 1984 Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361-375.
Weiss,D.J. & McBride,J.R. 1984 Bias and information of Bayesian adaptive testing. *Applied Psychological Measurement, 8*, 273-285.
許紅・繁桝算男 1990 項目反応理論と教授内容の階層的構造表現による問題項目の提示順序の最適化 日本教育工学雑誌 第14巻 73-80.