教育心理学研究, 2000, 48, 235—246

展　望

心理測定尺度のコンピュータ・テスト化に向けての最近の動向

廣　瀬　英　子[1]

本論文ではパーソナリティ，興味，社会的態度などの心理測定尺度のなかで，3件法，5件法などと呼ばれる多段階評定尺度について，測定方法に関する現在の動きを展望した。はじめに，現在コンピュータ・テスト化されている心理測定尺度にどのようなものがあるかを概観した。コンピュータ化の際には，紙筆式の場合との同等性を確認する必要がある。これまでの研究では，その多くは同等性の条件を満たすと判断されている。しかし，測定内容や測定対象者によっては，テスト形態が結果に影響する場合もあり，コンピュータ式の回答方法が適切かどうかを慎重に判断しなければならない。続いて，心理測定尺度に対する項目反応理論の現在の適用状況をまとめた。多段階評定尺度に適用できる多値型モデルにはいくつかの種類がある。これまでの研究で，因子分析で一次元性が確認されている尺度に対してこれらのモデルを適用することによって，さらに細かい項目分析を行うことができ，項目を精選して尺度の精度を高められることが確認されている。これらのことをふまえて，コンピュータ・テスト化された心理測定尺度に項目反応理論を適用して，それをさらに有効に活用する方法を検討した。適応型テストの形式，あるいは診断型テストの形式を取り入れること，また，紙筆式の発想にとらわれない新しいテストを考案することなどが今後の課題である。これまで行われてきた，既存の紙筆式の尺度をコンピュータ・テスト化する手続きは，将来的に心理測定をコンピュータ上で行うことを考えたときに必要ではあるものの，単調な作業であった。しかし，この段階でコンピュータ化を止めることなく，項目反応理論をいかした，内容面でも方法面でも新しい心理測定尺度づくりにつなげることが重要である。

キーワード：コンピュータ・テスト，項目反応理論，多値型モデル，コンピュータ適応型テスト

1　はじめに

心理測定の方法には観察法，面接法，質問紙法などの様々なかたちがある（塩見・金光・足立, 1982）。その中でも，質問紙法はもっとも広く用いられている測定方法である（鎌原・宮下・大野木・中澤, 1998）。質問紙法でよく使われる回答方法は，3件法・5件法などと呼ばれる，順に並んだ幾つかの段階の中から最適なものを選ぶ評定法である（本論文中では，この回答形式をとる尺度を多段階評定尺度と呼ぶことにする）。

過去3年間（1996年—1998年）の『教育心理学研究』誌に新しく発表された多段階評定尺度は64編にのぼる。その内訳は，5段階評定が32編，4段階評定が20編，3段階評定が5編，6段階評定が4編，7段階評定が3編となっている。今後も心理学の研究者が多段階評定尺度を研究用に作成・利用する機会は多いと予想される。本論文では，パーソナリティや社会的態度，職業興味などを測定する尺度，なかでも評定段階数が3つ以上の多段階評定尺度について，測定方法に関する現在の動きと課題を展望する[2]。

まず初めに，テストの実施方法が，質問紙法と呼ばれる所以でもある紙筆式（Paper and Pencil Testing）からコンピュータ式に移行される例が増えてきたことについて述べる。コンピュータの画面上に質問項目を呈示し，被験者にキーボードやマウスを通して回答を入力してもらい，結果を表示するまでの全ての過程をコンピュータを介して行う方法（Computer Based Testing: 以下，コンピュータ・テスト）はMMPI（Minnesota Multiphasic Personality Inventory）では20年以上前から利用されているが，近年のコンピュータの普及とともに，他の尺度でも試みられるようになってきている。そこで実際に現在どの程度コンピュータ・テスト化が進んでいるか，その進展状況をまとめる。

次に，コンピュータ・テスト化された心理測定尺度に項目反応理論を適用して，それをさらに有効に活用する方法について考えてみる。パーソナリティや社会的態度を測定する尺度に対して項目反応理論（Item

[1] 東京女子大学現代文化学部　〒167-8585 東京都杉並区善福寺2-6-1　eikohirose@mbk.sphere.ne.jp

[2] 文献資料は，PsycINFO及びERICを手がかりとして，主要心理学専門誌・専門書所載のものを中心とした。

TABLE 1 コンピュータ・テスト化されている主な多段階評定尺度

尺度	文献
Adjective Check List	Sanitioso & Reynolds(1992)
Balanced Inventory of Desirable Responding	Lautenschlager & Flaherty (1990) ; Miles & King (1998)
Beck Depression Inventory	George et al. (1992)
Edinburgh Postnatal Depression Scale	Glaze & Cox (1991)
Hamilton Depression Rating Scale	Kobak et al. (1990)
Harrington-O'Shea Career Decision-Making System	Kapes & Vansickle (1992)
Inventory of Work-Related Abilities	Staples & Luzzo (1999)
Minnesota Multiphasic Personality Inventory (MINI / MMPI-2)	Ben-Porath et al. (1989) ; Watson et al. (1992) ; Sukigara (1996) 村上 (1993) ; Pinsoneault (1996)
Rosenberg's Measure of Self-Esteem	Miles & King (1998)
State-Trait Anxiety Inventory	George et al. (1992)
Strong Interest Inventory	Hansen et al. (1997)
Unisex Edition of the ACT Interest Inventory	Staples & Luzzo (1999)
Yale-Brown Obsessive-Compulsive Scale	Rosenfeld et al. (1992)

Response Theory : Birnbaum[3], 1968 ; Lord, 1980) を適用して分析を加える研究は，以前よりも注目されるようになってきている。その状況を整理した後，項目反応理論の適用も含めた，今後の心理測定尺度のコンピュータ・テスト化の方向性を検討する。

2 コンピュータ・テスト化に関する研究

2.1. 紙筆式とコンピュータ式の同等性

TABLE 1 はコンピュータ・テスト化された心理測定尺度の中で，論文として発表されている主なものをまとめたものである。他に論文としては発表されていない尺度もかなりあると思われる。紙筆式をコンピュータ式に変えるにあたって注意する点は，Hofer & Green (1985) をはじめ American Psychological Association のコンピュータ・テストに関するガイドライン (APA, 1986) に示されている。そこでは，コンピュータ式で得られる結果が紙筆式の場合と同等(Equivalent)であることを示す必要性が強調されている。そのためには，両形式に回答した被験者の得点順位がほぼ対応していること，平均・分散・得点分布の形がほぼ等しいか，尺度変換すれば等しくなることが求められている。

職業適性検査についてのごく最近の研究に Staples & Luzzo (1999) がある。Holland の職業選択理論の6つのモデル環境に対する好みと，各々の職業分野で成功するために重要な能力を自己評定させる2つの尺度 (Unisex Edition of the ACT Interest Inventory, Inventory of Work-Relevant Abilities) は，紙筆式では仕事内容を文章のみで表現していたが，コンピュータ式 (CD-ROM) では，紙筆式と同じ説明文とともにその映像がテレビ画面に出るうえ，その説明文が音声でも流れるようになった。この工夫によって，被験者はよく知らない職業についても，そのイメージをつかむことができる。Staples & Luzzo (1999) では，各下位尺度ごとに紙筆式と CD-ROM 式の得点に高い相関がみられること，形式ごとに算出した下位尺度得点の平均値に，形式の違いによる差がほとんどみられないことなどの点から同等であるとして，等パーセンタイル法により基準の等化を行っている。

他にも Holland の理論に基づいた興味検査の中では，Harrington-O'Shea Career Decision-Making System での比較 (Kapes & Vansickle, 1992)，Strong Interest Inventory での比較 (Hansen, Neuman, Haverkamp & Lubinski, 1997) があり，紙筆式とコンピュータ式は基本的に同等と考えて差し支えないことが示されている。Kapes & Vansickle (1992) では，両形式に差がみられるかどうかを，再テストの結果も含めて分散分析によって確認している。

パーソナリティ測定尺度の中では Adjective Check List がコンピュータ化され，紙筆式との比較が行われている (Sanitioso & Reynolds, 1992)。コンピュータ式は評定カテゴリを“Yes”，“No”，“Don't Know”の3種類に増やした点で紙筆式と異なるが，得られたプロフィールはほとんど重なるものであった。抑うつ性の測定では，抑うつ症状を示す患者が Hamilton Depression Rating Scale にコンピュータで回答した場合，紙筆式の場合と同じような結果が示されることが確認されている (Kobak, Reynolds, Rosenfeld & Greist, 1990)。Yale-Brown Obsessive-Compulsive Scale は脅迫神経症の程度を評定する尺度であるが，患者がコンピュータに回答した場合，やはり紙筆式の場合と同様の結果が示された。そしてコンピュータ式も，脅迫神経症傾向があるかどうかの判別に使えることが確認さ

[3] Birnbaum (1968) の中では latent trait model と呼ばれている。

れている（Rosenfeld, Dar, Anderson, Kobak & Greist, 1992）。

MMPIに関しては，両形式の比較が幾つも行われている[4]。Watson,Thomas & Anderson（1992）は，それまでに行われている9つの比較研究のメタ分析を行い，K,D, Hy, Pd, Pa, Pt, Sc, Stの8尺度で，コンピュータ式の方が平均得点が低くなり，基準を別にすることを提言した。Sukigara（1996）はMMPIの日本語翻訳版で比較したが，逆にD, Pa, Pt, Sc尺度でコンピュータ式の方が平均得点が高くなった。一方，村上（1993）の，日本語版MMPIをさらに短縮・改訂したMINIでの比較では，D, Pd, Si, ?, F, SUS尺度で有意差があり，?尺度以外はコンピュータ式の方が平均得点が低くなっていた。Pinsoneault（1996）がMMPI-2について行った比較では，どの尺度においても有意差は見られなかった。

コンピュータ式が紙筆式と同等であるかどうかを確認する作業は，外国語の尺度の翻訳版を作るときの作業と似ている。このような比較研究は必要なことではあるが，そのためだけに大掛かりな実験を行うことは実際問題として非常に難しい。どの尺度について，また，どのような被験者集団において同等性の確認がなされているかを多くの研究者が知ることができるようにし，得られた結果を共有していく必要がある。

2.2. コンピュータ・テスト化の影響

コンピュータ・テストの形態は，被験者から肯定的に受け入れられていることが報告されている（Parshall, 1995）。しかし，測定する内容によっては，コンピュータ式であることが被験者の回答に直接影響を与え，本来の測定目的に障りがでることもありうる。例えば，コンピュータ不安を持つ被験者への影響が考えられる。George, Lankford & Wilson（1992）の研究では，コンピュータ・テスト形式のBeck Depression Inventory[5]（BDI）とState-Trait Anxiety Inventoryを受験した被験者の，あらかじめ紙筆式で確認されたコンピュータ不安の程度が，BDI得点とだけ高い相関を持つという結果が得られた。テストの間，コンピュータという苦手な刺激に注意が向き，それが抑うつ性の得点に反映したのではないかと考えられる。ただし，特性不安（Trait Anxiety）はともかく，状態不安（State Anxiety）とコンピュータ不安の相関が低いところに疑問が残る研究結果であり，今後のさらなる検討が必要である。

また，コンピュータ・テストを実施するには，被験者がコンピュータ機器の基本的な操作技術を持っていることが前提となるが，現時点では世代によってコンピュータ・リテラシーの程度に差があることは否めない。心理測定尺度をコンピュータ化するときは，測定対象者にとって，そして測定内容にとってそれが適切であるかどうかを慎重に判断しなくてはならない。

ただ，平成10年12月に告示された新学習指導要領により，平成14年度から『総合的な学習の時間』を中心に小学校段階からコンピュータに慣れ親しむことになる。コンピュータ・リテラシーの点では問題が少なくなる方向にある。

一方コンピュータ・テスト化は心理測定に好ましい効果ももたらしている。心理学の測定では社会的望ましさを意識して被験者が回答を歪めるかどうかが問題とされることがあるが（岩脇, 1973），Miles & King（1998）では，コンピュータ式で回答すると，紙筆式に回答するより社会的望ましさを考えての答えが減少し，正直な回答が増えることが示されている。しかし，この種の研究では，被験者が“社会的に望ましい”回答をする必要のある場面や状況を設定することが難しいためか，Lautenschlager & Flaherty（1990）では全く逆の結果が得られている。いずれにしても，Koch, Dodd & Fitzpatrick（1990）では，コンピュータに対して回答する方が正直に答えやすいという被験者からの評価が得られており，今後研究を重ねる価値があると思われる。

3 項目反応理論の適用

3.1. 心理測定尺度と項目反応理論

パーソナリティや社会的態度を測定する尺度に対して項目反応理論を用いて分析を加える研究は，学力テストに対する適用に比べると数少ないが，テスト理論の研究者によってかなり前から行われていた（Thissen, Steinberg, Pyszczynski & Greenberg, 1983）。項目反応理論では，同一の尺度に含まれる質問項目は，同じ一次元の心理特性を測定するものと仮定し，被験者をその想定された一次元の軸上のどこかに位置づけようとしている（芝, 1991；渡部, 1993；池田, 1994）。項目反応モデルは，被験者の各項目に対する反応パタンを利用して心理特性の尺度値を推定するために用いられる。また各項目の持つ特性値も推定することができ，被験者とは独立に，各項目がその想定された軸上に位置づけられる。心理測定尺度に項目反応理論を適用することにはいくつかの利点がある。

[4] 3段階の評定形式ではあるが“cannot say”への回答をなるべく少なくすることが求められている。

[5] 厳密にはここでいう多段階評定尺度ではない。

第1は，項目分析を掘り下げて行うことができる点である。心理測定尺度を作成する際には，因子分析が利用されることが少なくない。因子分析の結果，同一因子に含まれると判断された項目群に項目反応モデルを適用することは，一次元性の仮定に矛盾しない。項目反応モデルにはいくつかの種類があり，大きく分けると，“はい・いいえ”などの2段階のデータを扱う2値型モデル(Dichotomous Model)と“非常にあてはまる”から“全くあてはまらない”まで3段階以上のデータを扱う多値型モデル（Polytomous Model）がある。

多値型モデルを適用すれば，項目の持つ各評定段階(カテゴリ)ごとに特性曲線を描くことができる(Thissen & Steinberg, 1988)。項目反応カテゴリ特性曲線は，被験者の持つ心理特性の尺度値θと，各カテゴリに反応する確率との関係を表わしている。FIGURE 1には5つのカテゴリを持つ項目の例として2つの項目をあげている。項目A，Bとも，各カテゴリに対応するように5つの特性曲線が引かれているが，それは，あるθにおいてカテゴリ1，2，3，4，5を選択する確率を示している。項目Aでは，選択されたカテゴリの違いが，θのレベルの違いを明確に表わしている。しかし，項目Bでは，その関係が項目Aの場合ほど明瞭ではない。また項目Bのカテゴリ3はθのレベルにかかわらず選択される確率が低く，ひとつのカテゴリとして存在する意味が小さくなっている。カテゴリ特性曲線から見て問題のある項目は尺度から除外していくことによって，より良い尺度を構成することができる。

第2は，項目反応理論を尺度の等化に応用できることである。TOEFLのように複数機会の実施を必要とする学力テストにおいては，尺度の等化は大きな課題のひとつであり，様々な等化法が工夫されてきた。その中で，項目反応理論の応用は有力な手段となった(Kolen & Brennan [1995] 参照)。心理測定尺度においても，同様に，同一概念を測定する異なる尺度間の等化などに応用する可能性が考えられる。

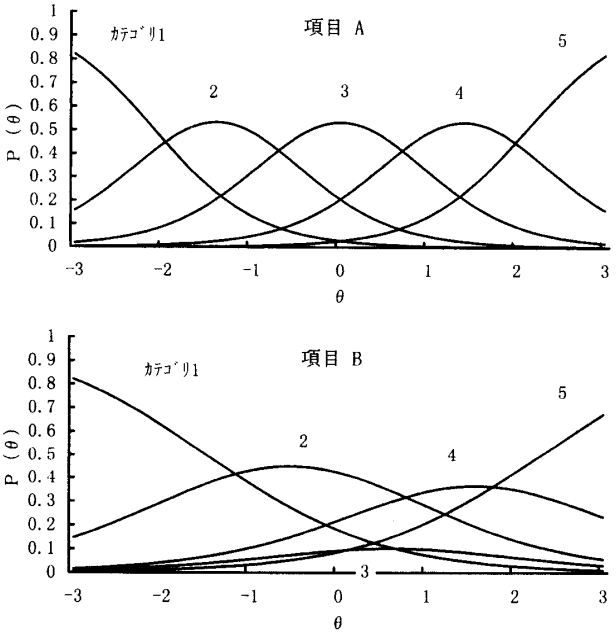

FIGURE 1 5段階評定項目のカテゴリ特性曲線の例

第3は，やはり項目分析の範疇に含まれることであるが，DIF (Differential Item Functioning) の検討を行えることである(Steinberg & Thissen, 1995)。DIFとは，項目への被験者の反応が，その尺度が測定しようとしている目的とは直接関係のない被験者所属集団の違いの影響を受けることである。そういう性質を持った項目は尺度から外すことが望ましい。特定の集団のデータを用いて描いた項目特性曲線を，他集団の場合と比較することによって，問題のある項目を明らかにすることができる。

3.2. 多値型モデルの適用

項目反応理論の多値型モデルは，理論的な考え方の違いにより，大きく2種類に分類されている(Thissen & Steinberg, 1986)。ひとつは，被験者がある順序づけられたカテゴリを超えるカテゴリを選択する確率を考えるモデルで，Samejima (1969) による Graded Response Model (段階反応モデル) が基盤であり，Muraki (1990) の Rating Scale Model もこれにあたる。もうひとつは，被験者がどの段階レベルに近いカテゴリを選択するか，その確率を考えるモデルで，Masters (1982) の Partial Credit Model (部分反応モデル)，Andrich (1978) の Rating Scale Model，Rost (1988) の Successive Intervals Model, Muraki(1992)の Generalized Partial Credit Model，Roberts & Laughlin (1996) の Graded Unfolding Model などが含まれる。

このようにモデルはいくつも発表されているが，既存の尺度にモデルを適用する応用研究には，段階反応モデルが使われている例が多い。モデル適用の際には必ず計算プログラムが必要となるが，商品化されたプログラムとしては，段階反応モデルと部分反応モデルが含まれた（ただし部分反応モデルについては利用者自身による若干の条件設定が必要）MULTILOG (Thissen, 1991) が唯一のものであったことが，その一因であろう（両モデルはDe Ayala [1993] にわかりやすく説明されている)。現在では PARSCALE (Muraki & Bock, 1997) も利用されるようになってきている。PARSCALE には Rating Scale Model (Muraki, 1990) と Generalized Partial Credit

Modelが含まれている。測定内容によるが，項目の持つ評定カテゴリ間の間隔が，同一尺度内に含まれる項目群全体で共通であると考えることができる場合はPARSCALEを，そうでない場合はMULTILOGを用いることになろう（どちらもAssessment System Corporation社, http://www.assess.com及びScientific Software International社, http://www. ssicentral.comを参照のこと。）

TABLE 2はパーソナリティや社会的態度などの多段階評定尺度の中で，実際に項目反応モデルが適用された例をまとめたものである。これらの研究では，項目反応理論は不適当な項目を見つけ出すのに極めて有効であること，そして尺度の精度を高めることに貢献出来ることが確認されている。また，以下のことも明らかにされている。

Baker, Zevon & Rounds (1994) は，気分 (mood) に関する問題をポジティブな言葉を使って訊ねる項目群とネガティブな言葉で訊ねる項目群に分けて別々にモデルを適用した。その結果，被験者の気分の状態(上下)によって，高い推定精度を得るに適した項目群が異なる様子が示された。尺度を構成する時に意図的に逆転項目を含めることがあるが，逆転しない場合と全く同じ測定ができているとは限らないことが示唆される。また，項目反応理論の多値型モデルでは，カテゴリごとにその特性曲線を描き，各カテゴリの働きを検討することができるので，評定カテゴリの並び順（程度の大小の向き）の影響がある場合に，従来の方法と比べて，それを検出することが容易になると考えられる。実際にChan (1991) は，被験者に呈示された評定カテゴリの並び順が回答に影響を与えることがあることを示している。評定尺度の作成には慎重な配慮が必要なことが示唆される。

ところで，カテゴリ間に順序性のある多段階評定尺度に対して段階反応モデルと部分反応モデルのどちらを使ったら良いかという問題もある。Maydeu-Olivares, Drasgow & Mead (1994) はシミュレーション実験で比較を行っている。5段階評定形式の項目を想定し，項目数は5項目から25項目まで，データ数は250から3,000までと変えていき，段階反応モデルとThissen & Steinberg (1986) のモデル（Masters [1982] の部分反応モデルの拡張形）に対する当てはまりの良さを比較したところ，この条件では両者に特に違いは見られないことを確認した。

3.3. 2値型モデルの適用

心理測定尺度の中でも“はい・いいえ”等の2段階で訊ねる形式のものについては，2値型モデルが適用されている。例えば，藤森 (1992) はTPI（東大式人格目録検査），MMPI，Y-G性格検査それぞれの社会的向性尺度に2パラメタ・ロジスティック（2PL）モデルを適用している。また，これまで見てきたように尺度が多段階評定の形式をとっているならば，項目反応理論の多値型モデルを適用するのが自然であろう。しかし，実際には，多段階評定のデータが得られているのにもかかわらず2値型モデルが適用されていることがある。Balasubramanian & Kamakura (1989) では6段階評定形式の消費者態度測定尺度に対して，また酒井・山口・久野 (1998) は5段階評定の価値志向性尺度に対して2 PLモデルを適用している。酒井他 (1998) の場合は，尺度内の項目が一次元階層性をなしているかどうかを知ることが主目的であるので，各項目の評定カテゴリが持つ特性まで明らかにする必要がなかったためと思われる。Balasubramanian & Kamakura (1989) の場合は2 PLモデルである必然性は見られない。研究の目的にもよるが，サンプル数が少ない場合や，特

TABLE 2 項目反応モデルが適用された主な多段階評定尺度

尺 度	文 献	使われたモデル	CAT[a)]
Attitude toward Capital Punishment Scale	Roberts & Laughlin (1996)	Graded Unfolding	
Attitude towards the Social Implications of Science	Foong & Lam (1991)	Graded Response	○
Attitude toward Women Scale	Dodd (1990) ; Dodd & De Ayala (1994) ; Chen et al. (1997)	Andrich's Rating Scale	○
Audit of Administrator Communication Scale	Dodd (1990) ; Dodd & De Ayala (1994)	Andrich's Rating Scale	○
	Koch(1983)	Graded Response	
Consumer Attitudes toward the Marketplace	Balasubramanian & Kamakura (1989)	2 Parameter Logistic	
General Social Surveys	Muraki (1990)	Muraki's Rating Scale	
価値志向性尺度	酒井・山口・久野(1998)	2 Parameter Logistic	
Mississippi Scale for Combat-Related Posttraumatic Stress Disorder	King et al. (1993)	Graded Response	
Mood Checklist	Baker et al. (1994)	Graded Response	
Personal Distress Scale	Chan (1991)	Graded Response	
Rosenberg's Self-Esteem Scale	Steinberg & Thissen (1995)	Graded Response	
State-Trait Anxiety Inventory	Steinberg (1994) ; Steinberg & Thissen (1995)	Graded Response	
Toronto Alexithymia Scale	Hendryx et al. (1992)	2 Parameter Logistic	

a) ○印はCAT（Computerized Adaptive Testing：コンピュータ適応型テスト）の形になっている尺度。CATについては4.1節参照。

定のカテゴリへの回答数が少なくて情報量の増加に役立たないというような場合は別として，せっかく段階数を増やして得た情報を分析の段階で不用意に潰すことは避けた方が良いと思われる。

4 今後の発展の方向性

ここまで，パーソナリティや社会的態度，職業興味等の心理測定尺度について，回答方法のコンピュータ化と項目反応理論の適用の2点を概観してきた。これらの工夫によって，コンピュータ・テストとしての利用可能性をどのように広げていくことができるか，その発展的方法を考えてみる。

4.1. 心理測定尺度のコンピュータ適応型テスト化

コンピュータ・テストを受ける被験者は，コンピュータに直接回答を入力するので，コンピュータ側でその回答を逐一分析し，用意されている項目プールの中から次に出題する項目として一番適切なものを選んで，画面に呈示することができる。この方法はコンピュータ適応型テスト（Computerized Adaptive Testing : CAT）と呼ばれている。適応型テストの発想は，今世紀初頭に作られたA.Binetの知能検査において，発達段階に応じて問題を呈示する方法にすでに示されている。それが，コンピュータの発達と結びついて，現在のCATが形作られた。CATには，ほとんどの場合，項目反応理論が取り入れられている。このように，項目反応理論は尺度構成における項目分析以外でも重要な役割を担っている。

CATの概要はいくつもの論文や専門書にまとめられている。Wainer, Dorans, Flaugher, Green, Mislevy, Steinberg & Thissen (1990) や，最近のものではWeiss (1995) がある。日本でも芝 (1991)，池田 (1994) などで解説がなされ，柴山・野口・芝・鎌原 (1987)，服部 (1989, 1990)，藤森 (1995)，永岡・植野 (1992) などの研究が行われている。CATに関するガイドラインとしてはAmerican Council on Eudcation (1995) があり，実用化に絡んだ諸問題についてはMills & Stocking (1996) にまとめられている。

学力テストに対するCATの応用研究は数多い（中村, 1993 ; Young, Shermis, Brutten & Perkins, 1996）。最近では音楽の聞き取りテストにも取り入れられている (Vispoel, Wang & Bleiler, 1997) ほか，医師の資格試験 (Mancall, Bashook & Dockery, 1996) や，適性検査 (Armed Services Vocational Aptitude Battery : Sands, Waters & McBride, 1997) 等での実用化も進んでいる。

パーソナリティ・テストについても，Steinberg & Thissen (1995) は，いずれCATの形式で行われるのが普通になるであろうと述べているが，実用例はまだ少ない。現段階ではコンピュータ・テストとしての利用もまだあまり進んでいないので，まず，コンピュータ式という形態がテストの作成者，実施者，被験者の三者に浸透していかないと，なかなかCAT化にまで進んでいかないであろう。今のところ多段階評定形式を用いる心理測定尺度の中でCAT化が試みられているのは，Audit of Administrator Communication Scaleと Attitude toward Women Scale (Dodd, 1990)，及び，シンガポールの中学生の科学に対する態度尺度 (Foong & Lam, 1991) のみである。このFoong & Lam (1991) では，全部で48項目用意されているうちの，3分の1の項目で十分に精度の高い態度特性の推定ができることが示されている。また，はじめから多段階評定形式をとる心理測定尺度ではないが，野口 (1995, 1999) は，鉄道事業の運転従業員に対する職務適性検査の中の速度検査項目をCAT化する際に，各項目への解答および反応時間情報から，各反応に対して1から4までのカテゴリ値のいずれかを与え，段階反応モデルを適用して尺度構成を行っている。

1995年までの項目反応理論の多値型モデルを用いたCATの主な研究については，Dodd, De Ayala & Koch (1995) 及び平井 (1995) にまとめられている。それによると，シミュレーション・データによる理論研究が多い（De Ayala,Dodd & Koch, 1992 ; Dodd, Koch & De Ayala, 1989, 1993 ; Koch & Dodd, 1989, 1995）。それまでの研究では被験者特性値の推定が最尤推定法 (MLE) で行われてきたことに対して，1995年以降の進展は，Chen, Hou, Fitzpatrick & Dodd (1997) が期待事後推定法 (EAP) を用いたことである。その中で，EAPの事前分布に正規分布を用いても一様分布を用いても，MLEと同様に精度の高い推定値が得られることが確認された。これらの理論研究で得られた知見をもとに，今後は多くの尺度のCAT化を進めてその効果を確認し，実際に使えるようにしていく必要がある。

4.2. 被験者特性の診断における利用

CATでは，被験者の特性値を最終的に導き出すが，テストの目的によっては，単に被験者の特性値が一定のラインを超えているか否かを判定すれば良い場合もある。そこで，Computerized Classification Testing (CCT) というテスト方法が考えられている。CCTのもっとも一般的な方法は，項目反応理論を用いた

CATと基本的には同じである。CATと異なる点は，合否の境界値をあらかじめ設定し，Sequential Probability Ratio Testing (Wald, 1947) の仮説検定の考え方に基づいて判別を行うことである(Reckase, 1983)。この方法では，個人の特性値を特定するところまで要求されていないので，その分全体として計算の負担が少なくなるという利点がある。この方法は，これまで学力テストにおいて用いられてきたが，心理測定尺度に対して利用することも可能である。Waller & Reise (1989) は，このCCTの考え方を心理診断法に応用し，適応型テストで被験者の誤差を含む推定値の信頼区間が診断の境界値を含まなくなった時点で項目の呈示を止めるという方法をとっている。そもそも心理測定尺度の中には，診断を目的として作られたものもある。それらを診断型のコンピュータ・テストに移行することは，ごく自然であり，また実用的意味のあることである。

Ben-Porath, Slutske & Butcher (1989) は診断形式のMMPIを作成した。彼らはMMPIの各下位尺度に一次元性を仮定するのは無理があるとして，項目反応理論を用いる代わりにCountdown Methodを用いた。これは個々の尺度について，診断に必要な情報が集まったところで(例えば，30項目の尺度で“はい”を1点，“いいえ”を0点として20点を境としたとき，“はい”が20個，または“いいえ”が10項目集まった時点で)，項目の呈示を止めるという方法である。項目反応理論を適用できない場合でも，このような近似的方法で診断式のコンピュータ・テスト化を進めていくことができる。

4.3. 新しいテストの考案

既存の紙筆式のテストをコンピュータ化する場合には両者の同等性が重視されることは前述の通りである。しかし，コンピュータ・テストが紙筆式テストと同じ機能しか持てないと考える必要はない。全く新しくテストを作成する場合であれば，自由な発想を取り入れることができるはずである。Kyllonen (1991) は，基準とする紙筆式テストが何もない白紙の状態から新しいコンピュータ・テストを開発した。そして，紙筆式の様式を踏襲することにこだわらず，コンピュータ・テストに最善の方法を考案することが大事であると述べている。コンピュータ式では，視聴覚的に優れた質問呈示が可能である。文字の大きさや音量は被験者に合わせて調節でき，答えるべき箇所を色で弁別するようにすれば回答ミスも防ぐことができる。これらの細かい工夫は学力テストのCATではすでに実施され始めているが，より効果的な呈示条件を特定する研究も必要である。

テストの構成や展開の方法についても，紙筆式では実現できなかった手法をコンピュータ・テストで実現することが望ましい。課題の呈示方法においては，その例として，マルチメディア(音声や動画)を取り入れた方法を考えることができる。また，回答方法においては，多肢選択形式をとる学力テストの場合であれば，ある質問項目に対する解答として，1回目に選んだ選択肢が誤答であった場合には，続けて次に正答と思われる選択肢を選び，その手続きを正答に達するまで行うという，コンピュータ・テストならではの方法が考えられる。廣瀬 (1999) は，このような解答方式を達成式と呼び，その方式で得られる解答に対して，多値型の項目反応モデルを適用することを提案している。そして，シミュレーション実験により，被験者の特性値の推定精度が向上することが確認されている。また，典型値を測定する心理測定尺度の場合であれば，例えば，ひととおりの回答を終えた後に，その被験者の回答として統計的に起こりにくい不自然な回答がなされている項目について，再度回答を確認する方法が考えられる。この方法がうまく機能すれば，いわゆる逸脱した回答パタンが得られる可能性を低めることによって，より高い精度で被験者の特性値を推定できることが期待される。このように，パーソナリティや態度の測定尺度についても，コンピュータの特徴をいかした回答方法・測定方法を考えていくことにより，尺度の利用価値が高まることが予想される。

5 まとめ

このように，心理測定尺度のコンピュータ化の状況を整理して気づかされることは，個々のテストの実施形態を紙筆式からコンピュータ化する取り組みと，項目反応理論を用いた理論研究とのつながりが必ずしも強くはないことである。中村 (1999) は，日本国内の項目反応理論を利用した応用研究がこのところ停滞していることを指摘しているが，理論研究で得られた知見が実際の心理測定場面にうまくいかされないことは，残念なことである。臨床場面において，コンピュータ化によって採点にかかる負担が減り利用しやすくなるという肯定的な意見 (Flowers, Booraem & Schwartz, 1993; Glaze & Cox, 1991) は少なくないのである。被験者にとっても，回答項目数が少なくなれば，回答の際の負担が減ることになる。

コンピュータ・テストを実施するには，そのためのソフトウェアを用意する必要がある。パーソナリ

ティ・テストをコンピュータ上で実施するためのプログラムとしては，例えば Windows 用の The Examiner と C-Quest (Assessment System Corporation 社)，MS-DOS 用の TESTAN がある(Shmelyov, 1996)。適応型形式のテストを行えるソフトウェアとしては，Windows 用の FastTEST Professional (Assessment System Corporation 社) が最近完成された。これは，マルチメディアを取り入れたテストも可能となっている。その他にも，研究者が必要に応じて自らの研究用に作成・使用しているソフトウェアは少なくないはずである。それらが広く公開されれば，テストのコンピュータ化はかなり進むのではないかと思われる。

本論文で扱ってきた尺度は，主として個人差の測定を目的とするものであった。集団としての傾向を調査するための質問紙(社会調査等)でも，多段階評定法はよく使われており，そこでもコンピュータ化は進められている。例えば，マーケティングの分野において CAT 形式の調査が推奨されているし (Singh, Rhoads & Howell, 1992)，大学教育においても学期末に学生が授業内容を評定する"授業評価"を紙筆式でなくコンピュータ式で行うことが試みられている (Cates, 1993)。今後，多方面で心理測定尺度のコンピュータ化が進展していくであろう。小・中学校の教育実践の中での活用も考えられる。新しく設置される『総合的な学習の時間』ではコンピュータを利用しての学習が本格化するが，そこでは生徒自身による自己評価が重視されるという。コンピュータ上で実施する多段階評定尺度のかたちは，ひとつの評価形態として採用されて良いのではないだろうか。

情報通信ネットワークの面では，van der Linden (1995)が指摘したように，コンピュータ・テスト化された尺度は将来的にコンピュータ・ネットワークを介して国際的に利用されていく可能性が十分にある。日本でも，高野 (1999) が WWW 版教師用 RCRT を開発して Web サイトに載せており，また前川・菊地 (1998)，菊地・前川 (1999) はインターネット上で実施できる適応型テストのシステムを開発している。既に個人の Web ページにアンケートを載せること自体は誰でも手軽に行うことができるようになっており，インターネット上でテストを実施する際に付随して生じる問題について，正しく対処できるように真剣に取り組む時期が来ているのではないだろうか。

最後に，テストを実際に使う立場の研究者と理論面を追求する立場の研究者の協力によって，今後，コンピュータ・テスト環境が一層充実し，ひとつの心理測定の方法として確立されることを望みたい。

引用文献

American Council on Education 1995 *Guidelines for computerized-adaptive test development and use in education.* Washington DC : Author.

American Psychological Association 1986 *Guidelines for computer-based tests and interpretations.* Washington DC : Author.

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

Baker, J. G., Zevon, M. A., & Rounds, J.B. 1994 Differences in positive and negative affect dimensions : Latent trait analysis. *Personality and Individual Differences*, **17**, 161—167.

Balasubramanian, S.K., & Kamakura, W.A. 1989 Measuring consumer attitudes toward the marketplace with tailored interviews. *Journal of Marketing Research*, **26**, 311—326.

Ben-Porath, Y.S., Slutske, W.S., & Butcher, J.N. 1989 A real-data simulation of computerized adaptive administration of the MMPI. *Psychological Assessment*, **1**, 18—22.

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*. Reading, MA : Addison-Wesley. Pp.396—479.

Cates, W.M. 1993 A small-scale comparison of the equivalence of paper-and-pencil and computerized versions of student end-of-course evaluations. *Computers in Human Behavior*, **9**, 401—409.

Chan, J.C. 1991 Response-order effects in Likert-type scales. *Educational and Psychological Measurement*, **51**, 531—540.

Chen, S., Hou, L., Fitzpatrick, S.J., & Dodd, B.G. 1997 The effect of population distribution and method of theta estimation on computerized adaptive testing (CAT) using the rating scale model. *Educational and Psychological Measurement*, **57**, 422—439.

De Ayala, R.J. 1993 An introduction to polytomous item response theory models.

Measurement and Evaluation in Counseling and Development, **25**, 172—189.

De Ayala, R.J., Dodd, B.G., & Koch, W.R. 1992 A comparison of the partial credit and graded response models in computerized adaptive testing. *Applied Measurement in Education*, **5**, 17—34.

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., & De Ayala, R.J. 1994 Item information as a function of threshold values in the rating scale model. In M. Wilson (Ed.), *Objective measurement : Theory into practice II*. Norwood, NJ : Ablex. Pp.299—315.

Dodd, B.G., De Ayala, R.J., & Koch, W.R. 1995 Computerized adaptive testing with polytomous items. *Applied Psychological Measurement*, **19**, 5—22.

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.

Flowers, J.V., Booraem, C.D., & Schwartz, B. 1993 Impact of computerized rapid assessment instruments on counselors and client outcome. *Computers in Human Services*, **10** (2), 9—18.

Foong, Y., & Lam, T. 1991 (April) 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.

藤森　進　1992　項目反応理論の社会的向性尺度への適用　岡山大学教育学部研究集録, **89**, 211—217.

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, **65**, 446—453.

George, C.E., Lankford, J.S., & Wilson, S.E. 1992 The effects of computerized versus paper-and-pencil administration on measures of negative affect. *Computers in Human Behavior*, 8, 203—209.

Glaze, R., & Cox, J.L. 1991 Validation of a computerised version of the 10-item (self-rating) Edinburgh Postnatal Depression scale. *Journal of Affective Disorders*, **22**, 73—77.

Hansen, J.C., Neuman, J.L., Haverkamp, B.E., & Lubinski, B. R. 1997 Comparison of user reaction to two methods of Strong Interest Inventory administration and report feedback. *Measurement and Evaluation in Counseling and Development*, **30**, 115—127.

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, **13**, 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, **38**, 445—454.

Hendryx, M.S., Haviland, M.G., Gibbons, R.D., & Clark, D.C. 1992 An application of item response theory to Alexithymia assessment among abstinent alcoholics. *Journal of Personality Assessment*, **58**, 506—515.

平井洋子　1995　適応型テストの現状—能力測定と診断　東京大学大学院教育学研究科紀要, **35**, 187—195.

廣瀬英子　1999　正答に至るまでの解答経路を用いた被験者特性値の推定　日本教育工学雑誌, **23**, 99—108.

Hofer, P.J., & Green, B.F. 1985 The challenge of competence and creativity in computerized psychological testing. *Journal of Consulting and Clinical Psychology*, **53**, 826—838.

池田　央　1994　現代テスト理論　朝倉書店

岩脇三良　1973　心理検査における反応の心理　日本文化科学社

鎌原雅彦・宮下一博・大野木裕明・中澤　潤　1998　心理学マニュアル　質問紙法　北大路書房

Kapes, J.T., & Vansickle, T.R. 1992 Comparing paper-pencil and computer-based versions of the Harrington-O'Shea Career Decision-Making System. *Measurement and Evaluation in Counseling and Development*, **25**, 5—13.

菊地賢一・前川眞一 1999 Webサーバーを用いたコンピュータ適応型テストシステムの開発 日本行動計量学会第27回大会 発表論文抄録集, 159.

King, D.W., King, L.A., Fairbank, J.A., Schlenger, W. E., & Surface, C. R. 1993 Enhancing the precision of the Mississippi Scale for Combat-Related Posttraumatic Stress Disorder : An application of item response theory. *Psychological Assessment*, **5**, 457—471.

Kobak, K.A., Reynolds, W.M., Rosenfeld, R., & Greist, J.H. 1990 Development and validation of a computer-administered version of the Hamilton Depression Rating Scale. *Psychological Assessment*, **2**, 56—63.

Koch, W.R. 1983 Likert scaling using the graded response latent trait model. *Applied Psychological Measurement*, **7**, 15—32.

Koch, W.R., & Dodd, B.G. 1989 An investigation of procedures for computerized adaptive testing using partial credit scoring. *Applied Measurement in Education*, **2**, 335—357.

Koch, W.R., & Dodd, B.G. 1995 An investigation of procedures for computerized adaptive testing using the successive intervals Rasch model. *Educational and Psychological Measurement*, **55**, 976—990.

Koch, W.R., Dodd, B.G., & Fitzpatrick, S.J. 1990 Computerized adaptive measurements of attitudes. *Measurement and Evaluation in Counseling and Development*, **23**, 20—30.

Kolen, M.J., & Brennan, R.L. 1995 *Test equating: Methods and practices*. New York : Springer-Verlag.

Kyllonen, P.C. 1991 Principles for creating a computerized test battery. *Intelligence*, **15**, 1—15.

Lautenschlager, G.J., & Flaherty, V.L. 1990 Computer administration of questions : More desirable or more social desirability? *Journal of Applied Psychology*, **75**, 310—314.

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

前川眞一・菊地賢一 1998 Webサーバーを用いたコンピュータ適応型テストの試み 日本行動計量学会第26回大会 発表論文抄録集, 191.

Mancall, E.L., Bashook, P.G., Dockery, J.L. (Eds.) 1996 *Computer-based examinations for board certification*. Evanston, IL : American Board of Medical Specialties.

Masters, G.N. 1982 A Rasch model for partial credit scoring. *Psychometrika*, **47**, 149—174.

Maydeu-Olivares, A., Drasgow, F., & Mead, A.D. 1994 Distinguishing among parametric item response models for polychotomous ordered data. *Applied Psychological Measurement*, **18**, 245—256.

Miles, E.W., & King, W.C., Jr. 1998 Gender and administration mode effects when pencil-and-paper personality tests are computerized. *Educational and Psychological Measurement*, **58**, 68—76.

Mills, C. N., & Stocking, M. L. 1996 Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education*, **9**, 287—304.

村上宣寛 1993 MINI性格検査の冊子方式とコンピュータ方式の違いについて 心理学研究, **64**, 279—283.

Muraki, E. 1990 Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement*, **14**, 59—71.

Muraki, E. 1992 A generalized partial credit model : Application of an EM algorithm. *Applied Psychological Measurement*, **16**, 159—176.

Muraki, E., & Bock, R.D. 1997 *PARSCALE*. Chicago, IL : Scientific Software International.

永岡慶三・植野真臣 1992 多元的適応型テストシステムのアルゴリズム 日本教育工学雑誌, **15**, 157—165.

中村知靖 1993 項目反応理論におけるパラメータ推定問題と適応型テストの開発 東京大学大学院教育学研究科博士論文

中村知靖 1999 測定・評価に関する研究の動向 教育心理学年報, **38**, 105—119.

野口裕之 1995 識別性検査A-1001の「知覚の速さ・正確さ」領域のIRT尺度化 名古屋大学教育学部紀要—教育心理学科—, **42**, 59—71.

野口裕之 1999 段階反応モデルによるIRT尺度化

渡辺直登・野口裕之 編著 組織心理測定論：項目反応理論のフロンティア 白桃書房 Pp.230—235.

Parshall, C.G. 1995 Practical issues in computer-based testing. *Journal of Instruction Delivery Systems*, **9** (3), 13—17.

Pinsoneault, T.B. 1996 Equivalency of computer-assisted and paper-and-pencil administered versions of the Minnesota Multiphasic Personality Inventory-2. *Computers in Human Behavior*, **12**, 291—300.

Reckase, M.D. 1983 A procedure for decision making using tailored testing. In D.J. Weiss (Ed.), *New horizons in testing : Latent trait test theory and computerized adaptive testing*. New York : Academic Press. Pp.237—255.

Roberts, J.S., & Laughlin, J.E. 1996 A unidimensional item response model for unfolding responses from a graded disagree-agree response scale. *Applied Psychological Measurement*, **20**, 231—255.

Rosenfeld, R., Dar, R., Anderson, D., Kobak, K.A., & Greist, J.H. 1992 A computer-administered version of the Yale-Brown Obsessive-Compulsive Scale. *Psychological Assessment*, **4**, 329—332.

Rost, J. 1988 Measuring attitudes with a threshold model drawing on a traditional scaling concept. *Applied Psychological Measurement*, **12**, 397—409.

酒井恵子・山口陽弘・久野雅樹 1998 価値志向性尺度における一次元的階層性の検討—項目反応理論の適用— 教育心理学研究, **46**, 153—162.

Samejima, F. 1969 Estimation of latent ability using a response pattern of graded scores. *Psychometrika, Monograph*, No.17.

Sands, W.A., Waters, B.K., & McBride, J.R. (Eds.) 1997 *Computerized adaptive testing : From inquiry to operation*. Washington, DC : American Psychological Association.

Sanitioso, R., & Reynolds, J.H. 1992 Comparability of standard and computerized administration of two personality questionnaires. *Personality and Individual Differences*, **13**, 899—907.

芝 祐順(編) 1991 項目反応理論 基礎と応用 東京大学出版会

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, **35**, 363—367.

塩見邦雄・金光義弘・足立明久(編) 1982 心理検査・測定ガイドブック ナカニシヤ出版

Shmelyov, A.G. 1996 TESTAN : An integrated modular system for personality assessment and test development on MS-DOS personal computers. *Behavior Research Methods, Instruments, and Computers*, **28**, 89—92.

Singh, J., Rhoads, G.K., & Howell, R.D. 1992 Adapting marketing surveys to individual respondents. *Journal of the Market Research Society*, **34**, 125—147.

Staples, J.G., & Luzzo, D.A. 1999 Measurement comparability of paper-and-pencil and multimedia vocational assessments. *ACT Research Report Series*, 99—1.

Steinberg, L. 1994 Context and serial-order effects in personality measurement : Limits on the generality of measuring changes the measure. *Journal of Personality and Social Psychology*, **66**, 341—349.

Steinberg, L., & Thissen, D. 1995 Item response theory in personality research. In P.E. Shrout, & S.T. Fiske, *Personality research, methods, and theory*. Hillsdale, NJ : Lawrence Erlbaum Associates. Pp.161—181.

Sukigara, M. 1996 Equivalence between computer and booklet administrations of the new Japanese version of the MMPI. *Educational and Psychological Measurement*, **56**, 570—584.

高野 明 1999 インターネットを利用した教育臨床サービスについての検討—WWW版教師用RCRTを通じて— 東京大学大学院教育学研究科修士論文

Thissen, D. 1991 *MULTILOG user's guide*. Chicago, IL : Scientific Software International.

Thissen, D., & Steinberg, L. 1986 A taxonomy of item response models. *Psychometrika*, **51**, 567—577.

Thissen, D., & Steinberg, L. 1988 Data analysis using item response theory. *Psychological Bul-*

letin, **104**, 385—395.

Thissen, D., Steinberg, L., Pyszczynski, T., & Greenberg, J. 1983 An item response theory for personality and attitude scales : Item analysis using restricted factor analysis. *Applied Psychological Measurement*, **7**, 211—226.

van der Linden, W.J. 1995 Advances in computer applications. In T. Oakland, & R.K. Hambleton (Eds.) *International perspectives on academic assessment: Evaluation in education and human services*. Boston, MA : Kluwer Academic Publishers. Pp.105—124.

Vispoel, W.P., Wang, T., & Bleiler, T. 1997 Computerized adaptive and fixed-item testing of music listening skill : A comparison of efficiency, precision, and concurrent validity. *Journal of Educational Measurement*, **34**, 43—63.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R.J., Steinberg, L, & Thissen, D. 1990 *Computerized adaptive testing : A primer*. Hillsdale, NJ : Lawrence Erlbaum Associates.

Wald, A. 1947 *Sequential analysis*. New York : John Wiley & Sons.

Waller, N.G., & Reise, S.P. 1989 Computerized adaptive personality assessment : An illustration with the Absorption Scale. *Journal of Personality and Social Psychology*, **57**, 1051—1058.

渡部 洋(編) 1993 心理検査法入門 福村出版

Watson, C.G., Thomas, D., & Anderson, P.E.D. 1992 Do computer-administered Minnesota Multiphasic Personality Inventories underestimate booklet-based scores ? *Journal of Clinical Psychology*, **48**, 744—748.

Weiss, D.J. 1995 Improving individual differences measurement with item response theory and computerized adaptive testing. In D.J.Lubinski, & R.V.Dawis (Eds.), *Assessing individual differences in human behavior : New concepts, methods, and findings*. Palo Alto, CA : Davies-Black Publishing. Pp.49—79.

Young, R., Shermis, M. D., Brutten, S.R., & Perkins, K. 1996 From conventional to computer-adaptive testing of ESL reading comprehension. *System*, **24** (1), 23—40.

謝 辞

本論文の執筆にあたりましては，東京大学大学院教育学研究科 渡部 洋教授，南風原朝和助教授に御指導を頂きました。心より感謝申し上げます。

(1999.7.13 受稿, 2000.1.18 受理)

Computerization of Psychological Testing : Review of Recent Studies

Eiko Ikeda Hirose (College of Culture and Communication, Tokyo Woman's Christian University) Japanese Journal of Educational Psychology, 2000, 48, 235—246

The present paper reviews recent studies of psychological testing methods, focusing especially on those with personality, attitude, and interest scales using a rating scale format. The style of administration of an increasing number of these scales is changing from paper-and-pencil testing (P&P) to computerized testing. For computerization, equivalence to paper-and-pencil testing must be shown. Many studies have determined that equivalence is satisified. However, for some examinees and some content, this difference in style of administration may affect the results. In those cases, whether computerized testing is appropriate must be considered. Psychological testing scales are also increasingly being applied to item response theory. Several polytomous models are suitable for rating scales. By applying the models for item analysis, more accurate and desirable scales be obtained. Some effective computerized testing methods that use item response theory are discussed. In future research, computerized adaptive testing and computerized classification testing must be examined for these scales. Innovative testing methods that are not based on established paper-and-pencil testing ideas are also needed.

Key Words : computer-based testing, item response theory, polytomous model, computerized adaptive testing, paper-and-pencil testing