

測定・評価部門

評価・測定と認知心理学

椎名 乾 平

(早稲田大学教育学部)

はじめに

著者は評価・測定のいわゆる専門家ではない(ことにしておく)。数式の読める,あるいは場合によってはちょっとした数式を捏造できる認知心理学者であるという自己評価をしている。そこで「いわゆる」専門家とはひと味違う評価をするであろうし,またそうでなくてはならないと考えている。どのようにひと味違うのか?ひとつにはおそらく一人の認知心理学者からみたこの分野のやややぶにらみの論評を行うことになるだろう。また過去1年間の論文を材料に,論文の著者達に叱られそうな,あるいは「研究の文脈を無視した言いがかり」とでも言われてしまいそうな,そのような論評をくわえることになってしまうかもしれない。そこで本稿で取り上げることになる論文・発表の著者諸賢に深くお詫びしてから論を進めたいと思う。

最近の認知心理学の隆盛により,知的構造・情報処理プロセスに対する我々の知識は著しく増大したわけであるが,その成果が現在のところ評価・測定の分野にあまり影響を与えていないとすれば奇妙なことと言わなければならぬ。ところがこの種の奇妙な状況は昨今の現象ではないようである。例えばどの教育心理学の教科書にも Piaget の発達段階説が出ているし,また多くの教科書には知能テストの話も出ている。ところがよく考えてみるとこの両者は相いれない点が多い。そもそも発達段階説というのは各段階の質的相違を重要視する立場であり,量的発想になじまない(この辺の議論については, Green, Ford, & Flamer, 1971参照)。この両立しない二つの発想が仲良く(あるいは互いを黙殺しながら),共存してきたのは考えると不思議な現象である。もちろん色々なアプローチがあって良いという意見は十分理解できる。しかし,その場合にはアプローチの相互関係の理解,あるいはどこをどのように分業しているのかについての理解,が不可欠である。

1. 表現 (Representation) と測定

1.1 表現

認知心理学の主流派は表現 (representation) という概念を極めて重要視する。表現とは何かについてはこれも

様々な説があり一筋縄ではいかないが,

表現されるべき対象→何らかの形式的体系
という関係づけ (写像) をおこなうものとされる (Palmer, 1978)。

例えば10個の単語の意味的類似性の評定データを MDS にかけて,空間表現を得たとしよう。この場合10個の単語(の意味関係)が「表現されるべき対象」であり,(意味)空間が「形式的体系」である。導出された空間に対しては,2種類の解釈が成り立ちうる。1つは,この空間が心理学的に妥当な内的表現であると解釈する立場であり,ほとんど同型な実体が心内に存在すると考えることになる(最近では「脳内表現」という用語さえ用いられる)。もう1つの立場はデータの情報を縮約して見やすい形に変形したというものである。この解釈のもとでは,表現とは情報の変換であり,より使いやすく,よりわかりよい形にデータを加工するものである。したがって同型の実体が心的に存在する必要はない。どちらの場合でも表現は情報の欠落を含意するが,そこはあえて問わないことになる。言うまでもなく形式的体系には今述べた空間表現以外にも様々な可能性がある。認知心理学では「命題」のようなものを形式的体系と考えるのが主流派であるから (Pylyshyn, 1984), 認知心理学と量的測定論というのは実はあまりしっくりしないのである。

1.2 表現としての尺度化・数量化

テストとは人間の様々な特性の表現を作る作業なのだと解釈できる。例えば,100点満点の期末テストを行いその合計点を求める先生は,

個々の生徒→0から100までの整数

という表現を行っていることになる。この作業によって「個性豊かな」個人やその正答・誤答パターンは単一の数字になってしまい多くの情報が失われるが,そのかわり見通しのよさと扱い易さが生まれてくるわけである。また数字の大小は常に比較可能なので,数字によって表現された個人は比較可能,さらに順序づけ可能となるのである。従って,選抜や相対評価には適した表現になるのである。いわゆる潜在特性モデルもテスト項目や人の representation を作るものだと解釈できる(認知心理学的解釈)。すなわち IRT(項目反応理論)はテスト項目を項目特

性曲線に、人を実数(潜在尺度値)に対応づけるものといえる。また因子分析は、人をベクトルに対応づけ、項目もベクトルに対応づけるものと言えよう。

項目→関数空間

IRT:

人→ $[-\infty, \infty]$

項目→ベクトル

因子分析:

人→ベクトル

繰り返しになるがここで重要なのは、「対応づけ」および「表現モデル」という概念である。表現モデルとは何らかの形式的体系であり、人を表現モデル内に「対応づける」(「位置づける」)ことによって人の特性を表現するわけである。しばしばテストは潜在特性を測定するものと想定されており、潜在特性の表現モデルとしては先に述べたとおり実数軸やベクトルを想定するのが普通である。しかしながら、テストによる測定を

人→形式的体系

と一般化して考えるならば、表現モデルには様々な可能性が存在する。例えば形式的体系として、ルールの集合(Siegler, 1976), 多次元距離空間(MDS), 樹状図, プロダクション(Siegler, 1981)等を用いる可能性が存在する。もちろんこういうものが「測定」と言えるかどうかは議論が分れるところであろう。

1.3 「尺度」というものの実体性

テストの目的は、表現をつくることを通じての、人間の情報処理プロセスの性能・特性の評価と言えよう。そこで

「認知プロセスを良く言えば大局的に、悪く言えば粗雑に表現すれば尺度が生じてくる」、あるいは

「認知心理学はミクロ、テスト理論はマクロ」

というのが私の考え方である。1つたとえ話をしたい。エンコ(自動車が動かない)という事態を考えてほしい。自動車が動かなくなるのはもちろん確率的な問題である。ガソリンがない、オイルが切れている、バッテリーがあがっている、故障している。そして故障の場合には部品や部品間の運動関係が壊れていたりする。なんにしても、なんらかの原因があってその結果として、自動車が動かなくなるわけである。原因となる不都合は確率的に生起すると考えてよいので、自動車が動くか動かないかも確率的問題と考えてよい。さらに一台一台の自動車にも個性があり、様々な条件で動かなくなる確率も異なるはずである。そこで、自動車の「可動性」という次元尺度を考えて、この尺度上に各自動車を位置づけようという発想がでてくるかもしれない。

ところで世の中には総合指標というものが存在する(例えば、物価指数)。この種の指数は、複雑な対象の概括的な利便性の高い総合評価値を与えてくれる。しかしあくまで目安であって、物価指数が(潜在的)原因となって、土地の値段や、タクシー料金や、食品の値段が決まるといっては本末転倒というべきであろう。

さてさきほどの自動車の「可動性」という尺度は果たして実体なのだろうか?それとも物価指数のような「便利な指標」なのだろうか?私は実体ではなくてあくまで指標であると考えている。可動性を現実には左右している要因は、燃料切れや、故障等の物理・化学的な過程であるからである。

人がある項目に誤答するのを、自動車の「エンコ」のように見なすことは可能であろう(Van Lehn, 1990)。すなわち、意欲がない(燃料切れ)、知識を想起できない(バッテリー切れ)、解法の適用が間違っている(故障)等々の原因が確率的に生起して誤答をもたらすと考えられるわけである。これこそが問題の本質に迫る正統的な道なのかもしれない。また、人が性格テストの項目に異なった反応をするのを、様々な条件下における自動車の「走り」の特性のように捉えるのも同様に可能であろう。

一次元連続体上の尺度というものは実体ではなく、人間(特に研究者)の認知プロセスが醸し出す幻なのかもしれない。もちろん人間みんなが納得する幻は(錯視のように)「心理学的事実」であるから、その意味で心理学的には妥当性があることになる。しかし人間の「尺度」や「テスト」に対する認知が変わればやはり幻であったと評価が変わる可能性がある(最近の偏差値排斥運動には評価に対する認知を変えるべきだという発想が確かに含まれていると思う)。

さて以上の議論が正しいとしたら、次元尺度のようなものを仮定するのはまったくの誤りなのだろうか?そうとはいえないと思う。1つには、精神物理学での感覚強度のように連続量の心的変数は確かに存在するからである。また抽象性の高い測定対象(不安、価値観、性格特性、等々)に対しては概括的な尺度のほうがあてはまりがよいし納得できるような気がする。さらに先ほどの例にもどれば、そもそも自動車の走るメカニズムがまったく不明であるのなら、概括的な構成概念を立てるような方法論が必要だし、またそれしか道はないと思われるからである。すなわち対象の細かい性質がよくわからないので、概括的な議論しかできないしまたそれでいいわけである。例えば「性格テストの項目に人間が反応するときどのような認知的処理過程が存在するか」というような問題を扱った研究は(私の知る限り)ほとんどない。

抽象的構成概念の間の関係をあつかうと、現実世界との関係が希薄になるので砂上の楼閣を築く懸念がでてく

るわけであるが、この危険性の存在はすでに常識になっていようし、因子分析は構成概念自動合成装置であるという嫌みもよく聞く。何が存在するかは人間の主観的信念によるところも多い。世の中には明らかに実体であるが、しかしながら直接的に測定できないものがある(例えば、恒星までの距離、湖に住む魚の数)。心的能力や潜在能力をこの種の直接測定不能だが確かに存在するものと「信じれば」、それは存在することになってしまう。しかし自動車の可動性尺度のようなものと考えれば概括指標(あるいは仮説的構成概念、あるいは仮定された実体)にすぎなくなり、いつの日か実証されるか反証されるか改訂されるべき対象となる。これは「テスト」や「尺度」に対するメンタルモデルの問題である。

正答誤答のあるテストに話を限定すれば、あるテスト項目に誤答するのは潜在能力に規定される確率が存在するからだと考えのではなく、項目に対する処理プロセスの不調、スキルやストラテジーの適用失敗の結果ととらえることになる。これは私だけの意見ではないらしい。例えば、Educational Testing Service の Mislevy (1993) は、「従来の「テスト理論」というのは19世紀の心理学に20世紀の統計学を応用したものにすぎず、単一の連続体上での仮想的「能力」について議論しているため、知識や問題解決に含まれるコンポーネント、スキル、ストラテジー、情報処理過程等の性質を明らかにするのは難しいだろう」と主張している。全く同感である。ただしもちろん従来のテスト理論が誤っているとか時代遅れであるというわけではなくて、認知心理学の成果を取り入れたテスト理論はテスト理論を選抜・予測問題から教授過程問題に適用可能にするために、望ましい発展・新しい展望を与えると期待できるだろうということである。

認知モデルや教授モデルを包含するテスト理論の米国での動向は Frederiksen, Mislevy, & Bejar (1993) に詳しい。この本は「新世代のテスト」という題名からもわかるように、テスト理論のパラダイムシフトを要求するもので、将来この分野の分岐点であったと評価されるものになるかもしれない。この他にも発達モデルを組み込んだ Falmagne, Koppen, Villano, Doignon, & Johanneson (1990) や Falmagne (1993) が注目に値する。最近明らかになってきた米国の動向はまさに、「受検者の知識状態」を測定可能ならしめるようなテスト理論の構築に向かっており、これは認知心理学の発展という事実を鑑み一つの歴史的必然であろうと思われる(もちろん冷やかな見解も多いし、線香花火に終わる可能性もあるが)。

1.3.1 尺度連続体の心理学的意味

一次元連続体上の尺度が心理学者(あるいはもっと広く世間の人々)に好まれるのはなぜだろうか?一つには現在に

至るまでの歴史的な経緯があるだろう。また実数というものが身近で他の選択肢がほとんどないというのも重要な要因であろう。また多くの場合尺度の使用目的が、選抜や大局的順序づけであると言う点も見逃すことはできない。

さらに立ち入った認知心理学的な理由も考えられる。すなわち人間には

- 1: 行動には何らかの原因があると考える傾向、
- 2: 情報を簡略化しようとする傾向、
- 3: 線形加算型の情報処理を行う傾向、

があるようだ。(1, 2の根拠はたくさんありすぎるので述べない。3の根拠は Brehmer (1979) や Dawes & Corrigan (1974) 参照)。とすればテストの合計を求めて、それを潜在能力と見なす傾向の根拠をこの辺に求めるのは可能であろう。Rasch モデル以外の IRT でも「潜在能力が高い人は、項目に正答する確率が高い」のであるから、

$$\text{能力 } \theta_i > \text{能力 } \theta_j$$

ならば

$$P(\text{項目 } k \text{ に正答} | \theta_i) \geq P(\text{項目 } k \text{ に正答} | \theta_j)$$

という関係が必ず成立し、よって正答数の多い人が高い潜在能力値を与えられる傾向が存在し、合計点との相関はかなり高くなり、概括指標的な色彩を帯びてくるのである。

1.4 新しい学力観と測定

学校教育の現場では指導要領改訂と「新しい学力観」に関して未だに混乱があるようである。確かに「新しい学力観」で推奨される、「関心・意欲・態度」や、「思考力・判断力」が簡単に測定・評価できるものとは思われない。また到達度評価や形成的評価もかけ声ばかりで、それほどたやすく実行できるとは思えない。さらに一次元尺度である偏差値のようなものは、使わないようにするというのが、最近の学校教育で流行の方針のようである(もっと露骨に言えば「能力」はなるべく評価するなということだろうか?)。というわけで、評価・測定の専門家には様々な社会的要請があるはずなのだが、この種の研究はあまり活発ではないように思える。北尾(1995)の本には様々なアプローチが列挙されているが、統一的な見解に至ることははたして可能なのだろうかという気がしてならない。

私見であるが、「思考力の測定」をそれほどおそれる必要はないのではないだろうか。ある課題が与えられた時、生徒・児童が思いつく解法は何パターンかに限定されることも多いはずである。どの解答パターンを用いようとするかで、どのような知識を用いてどのような解決をしようとしているのかある程度は推測可能だと思われる(少なくともベテラン教師は授業中にこの種の推測を行っているも

のと考えられる, 吉崎(1991))。もしそうなら課題に対する事前知識や学習者モデルを積極的にテスト理論内で利用することによって評価・測定問題も展望が開けてくるのではないだろうか? そもそも「思考力」の評価というからおかしくなる。現実にも求められているのは「思考」の評価であろう。ある特定の課題においてどう思考したかという内的思考状態をある程度トレースできればよいはずで, 抽象的思考力のようなものを求めよという要求はなされていないと思われる。

Piaget の研究や, ある種の課題分析は鋭い洞察に満ちているが, この洞察を積極的に測定モデルに組み込むといった試みは従来あまり存在しなかったように思える。先に述べたような表現モデルの柔軟化により知識・概念構造のより生々しい姿を測定という形で明らかにできるのではないか。さらに被検者がどのような問題解決スキルあるいはストラテジーを用いているかを, またどのスキルに欠陥があるかを明示することにより, 到達度評価あるいは形成的評価のために新しい道具を与えることができるのではないかと期待している。

2. 1994年7月から1995年6月にかけての主な進展

以上述べたことを念頭に置きながら各研究について考えたいと思う。論評するのは, 基本的に書籍・論文となって発表された研究である(良い学会発表は後日論文となって発表されるだろうから)。

2.1 尺度化・テスト作成

たくさんの尺度化研究が報告されており, その内容は多岐にわたっている。このセクション(2.1)で紹介する論文で, 「教育心理学研究」, 「心理学研究」にこの期間に掲載された論文の1割強を占めており, 一大勢力と言えるであろう。

2.1.1 不安尺度

萩生田(1995)は新しい不安尺度の構成をし, かなり充実した妥当性の検討を行っている。藤井(1994)は数学不安尺度を作成し, さらに数学不安の要因について言及している。不安のような対象に対しては認知心理学は今のところ無力である。

2.1.2 行動評定

この種の尺度化の意図は, 尺度を作成しながら現象を整理しよりよく理解しようということであろう。菅原ら(1994)は乳幼児期にみられる行動特徴を調べるための日本語 RITQ および TTS の尺度化を行った。中野(1995)は女性を対象とする Type A 行動パターン質問表を作成した。

2.1.3 小論文・作文

平(1995)の作文能力評価の研究は難しい対象を扱った

意欲的な研究である。課題は絵を見せてお話を作らせるものである(TAT や CAT の研究と関係がありそうだ)。それをベテラン評定者群に評定させ高い信頼性を得ている。平井・渡部(1994)は1つの小論文を多数の評定者が評価する状況下において, 評定カテゴリー数(2, 3, 4, 5)が評定の信頼性に与える影響を調べた。結果として5カテゴリー程度で素点(100点満点)と同様の信頼性が得られたと報告している。

今述べた2つの研究では, 共に小論文(作文)を「人」, 評定者を「テスト項目」のように見なしているが, この解釈の元では, 人間の作文能力ではなく, 小論文の「良い評定値を得る能力」の議論をすることになるので注意が必要であろう。すなわち人の評価ではなくて, より直接的には作品の評価なのであり, 被検者の「作文能力評価」ではなくベテラン評定者の「作文評価能力」の研究になっていないのか疑問が残る。一般に評定者をテスト項目と見なすのには注意が必要だと思う。例えば, 同じ評定者群がある被検者群の音楽演奏を評定し, 他の被検者群の作文の評価をするような状況を考えていただきたい。音楽演奏と作文の良さを同じ尺度上で評価できてしまう(これが利点となる場合もあるかもしれないが。ただしその場合には作品・演奏の「芸術としての良さ」のような構成概念を考えねばなるまい)。

2.1.4 その他

苧坂・苧坂(1994)は「読み」の情報処理プロセスに重要な働きをすると仮定されている, ワーキングメモリーの容量を測定するテストを作成した。認知プロセスの中の1つのコンポーネントに注目しその性質を調べようというアイディアは極めて新鮮に映る。ただしテスト理論の立場から言えば, 信頼性への言及がないなど不満も残る。

下山(1995)の「男子大学生の無気力について」の研究は, 「測定尺度の設定」, 「要因間関係の吟味」, 「共分散構造分析による因果モデル」という理想的3段階を踏んだ緻密な力作である。その結論について私はコメントする力はないが, 方法論的には教育心理学の研究法の手本となるのではないか。

柏木・山田(1995)は正準相関分析を用い, 性格特性5因子モデル(FFM)と内田クレペリン検査の各指標との関係を検討し, 内田クレペリン検査による性格特性評価の可能性についてポジティブな結果を得た。

平らの研究(1995)は語彙理解テスト・漢字読み取りテストの尺度化の研究である。大量の項目・大量の被検者を用い最新のテスト理論を用いて尺度化したもので労作の一語につきる。実際の教育場面で診断的テストとして有用なことも多いだろう。ところでこのテストで測定さ

れる「能力」は実在しているのだろうか？このテストの一つの使い方として、中級以上の日本語を学習した外国人に実施するということがあろう。ある教養ある外国人がこのテストを受けてあまり良い点数でなかったとする。その時この外国人は語彙理解「能力」が低いと言ってよいのだろうか？このテストの論理から言えばもちろん良いはずである。しかし「能力」というよりも学習量や学習時間を強く反映する尺度である可能性はないのだろうか？

その外に、木内 (1995) は独立・相互依存的自己理解尺度を、矢口 (1995) は音楽作品の感情価測定尺度を、濱口 (1994) は児童用主張性尺度を、大谷・桜井 (1995) は多次元完全主義尺度を報告している。

2.2 テスト理論とその周辺

池田 (1995) のモノグラフは古典的テスト理論から項目反応理論 (池田は項目応答理論と呼んでいる) までの話題を簡潔にまとめたもので、テストにかかわる人の必読書と考えられる。古典的テスト理論の歴史的展開や、付属プログラムによるシミュレーションの部分は特に興味深かった。

Ogasawara (1995) は項目反応理論における個人の尺度値が正規分布をし、さらにその平均と標準偏差が、他の変数の関数となっているモデルを提案した。応用例として、尺度値の分布が年齢の関数となっているモデルを20才から54才までの知能テストの結果に当てはめた結果が示されている。結果として Binet のテストのような年齢の軸上でのモデルになっているのが興味深い。

2.2.1 テストにおけるバイアス

被検者に与える不安感の心理的影響や、被検者のとる解答ストラテジーは、テスト結果に大きな影響を与えるはずで、数式をいじってわずかな推定精度の差を云々するような研究よりよほど本質的な問題かもしれない。

藤森・中野 (1994) の研究は、テストの教示内容を変えて回答行動の変化を探っている。「正答数を得点とする」あるいは「正答数引く誤答数を得点とする」という異なる教示を与えると、後者において、一般に無答数が増加し、解答数の多い積極的な被検者と解答数の少ない消極的な被検者があらわれ、さらに無答の発生傾向に個人差が存在することが示唆された。無答が多くなると IRT による個人能力推定値が不安定になるために後者のような教示は避けるべきだと結論している。解答するか無答ですますかは、自分の解答の正答確率の自己評価を含むメタ認知的意思決定過程であり、さらに正答確率の自己評価値は能力の関数になっていると思われるため、この種の研究では「解答するかどうか」をいかに決定するかをモデル化する必要があると思われる。

長谷川 (1995ab) は、自己評価的意識尺度 (項目例、「私は自分自身に自信を持っています」、「私は時々自分自身がいやになることがあります」、等々) での、評価性の低い項目において、自己評価の低い被検者群が1か月の後にプラスの方向に自己評価を変化させる傾向にあるという再検査効果を見いだした。もし多くのテストで同様の現象が起これば再検査信頼性の推定手続を再考しなければならないかもしれない。

孫・井上 (1995)、井上・孫 (1995) は差異項目機能に注目している。差異項目機能とは、下位集団によって正答率が異なる現象であり、特に人種問題とのからみで米国において注目されてきた。下位集団が存在するときに正答率が異なるとしてもそれは当たり前のものであり、これが問題となるのはテスト作成者が意図しないような差異が生じる場合である。異なる解法過程を用いる集団の存在を探る道具としておもしろいかもしいかなという主張は理解できるが、あえて差異項目機能という名の元で研究する必要があるのかないかよくわからない。

2.2.2 適応型テスト

植野・大西・繁樹 (1994) の研究が光る。ベイジアンネットワークを用いて教材についての領域構造知識をモデル化し、その上で適応型テストを用いる試みである。領域知識構造のフォーマットやベイジアンネットワークのような確率モデルの妥当性については異論があるかもしれないが、これからの研究領域として最も興味深い方向であろう。

適応型テストにおいては推測効率の観点から正答率 .5程度の項目が使用されるのが普通であるが、これはテスト受検者の「心理」を無視した方策かもしれない。例えば正答率 .5の問題を連続して与えられると、モチベーションを失う人もいるかもしれない。不安を感じずにテストを受けられるとしたら、正答率はどのくらいであるべきなのだろうか？藤森 (1995) は項目反応理論の文脈で正答率 .8前後が受検者にとって最適な困難度であるという結果を得ている。

2.2.3 大学入試

大学入試の妥当性は大きな問題であるが、大学の成績と入試成績の相関は、入試で選抜が行われるために原理的には推定不可能である。繁樹・萩生田 (1995) はギブスサンプリング法によって欠損値を推測しながら相関係数を推定する大変興味深い試みを行った。選抜効果を消し去るブレイクスルーになるかもしれない。田栗・橋本 (1995) は大学の教養課程の成績 y を、1次試験成績 x_1 、2次試験成績 x_2 から予測するための、最適な重みを ω_1 、 ω_2 を与える手法を提案した。ただし通常の重回帰ではなく

$$\omega_1 + \omega_2 = 1$$

$$\omega_1, \omega_2 \geq 0$$

という制約がついているところがミソである。

2.3 統計学と多変量解析

制約付き主成分分析では主成分分析における変数や観測個体に付加的な外部情報が与えられているときに、これらの情報を組み込んだ形で分析を行う。また、この手法は主成分分析はもちろんのこと、数量化3類、冗長性分析、計量的多次元尺度法、ベクトル選好モデル等々を特殊例として包含する。高根(1995)のモノグラフは高度な内容を扱った重厚なものであり、理論的明快さと応用可能性の大きさから最近発行されたモノグラフの中でももっとも重要なものの一つであるのは確かである。柳井(1994)のモノグラフは多変量解析の概説書であるが、応用例を重視している点と、最近の話題を積極的にとりいれている点で、初学者からプロレベルまで役にたつ本といえよう。武藤の大著(1995)(600ページ以上ある)は初学者を対象にしたものだが、通常のデータ解析での疑問点はすべてこの本で解決できるだろうという意味で、またすべて2ページ読み切りであるという点で、あまり統計に詳しくない学部高学年、あるいは大学院生に勧めたい本である(高価なのがやや難点だが)。

統計モデルの当てはまりを評価するのに、AICやBIC等の情報量基準がよく使用される。萩生田・繁樹(1995)はBICを改訂したNBIC基準を提案した。論文になって公刊されるのが楽しみである。

2.3.1 共分散構造分析

共分散構造分析は完全にポピュラーな手法として認められてきており、これからもますます適用例が増えると考えられる。しかし理論的にはまだ発展の余地も多いようであり(狩野(企画), 1995)、同値パスモデルと識別性の問題について重要な進展がみられた。同値パスモデルの存在とは、同一の適合度を持つモデルが複数存在することであり、さらに仮定される因果関係が逆転する場合さえありうるので共分散構造分析で何らかの因果推論する際には極めて重要な問題である。前川(1995)、Mayekawa(1994)は、パス行列と分散共分散行列の両立性を調べる方法と、同値パスモデルを生成する方法について新しい結果を導き出している。識別性の問題とは、あるモデル下での解が一意に定まるかという問題である。Toyoda(1994)は識別性を判別する条件を導き出している。

2.4 ニューロとファジィ

ニューラルネットワーク理論やファジィ理論を用いた研究も本格的なものが現れつつある(ニューラルネットワーク理論の火つけ役は、高名な認知科学者Rumelhartであるのを、またファジィ理論の創始者Zadehは心理学を絶えず念頭に置いてきたのを指摘し

ておきたい)。まず目につくのはニューラルネットワークを非線形回帰的な統計手法として使用しようとする方向である(麻生 他, 1995)。工学分野ではこの使い方は前からよく知られていた(例えば、パーセプトロンは線形判別分析と同等である)。ニューラルネットワークは本来は学習のモデルなので、教育心理学の他の分野の研究とも関係が深いはずである。

またファジィ制御を用いた評価モデル(清水, 1995)や、この方向のモデルを集めた原著が発行された(山下, 1995。ただしコメントは来年度となる)。清水のモデルは理論的には興味深い、モデルの当てはまりの良さや、信頼性、妥当性にさらなる検討を加える必要があるのかもしれない。

3. おわりに

以上、本年度は従来とはやや異なる視点から論評を行ってきた。やや意地悪な論評もあったかもしれないがご容赦いただきたい。

テストというものに対するメンタルモデルを変えるべき時なのかもしれない。少なくともテストに対する様々な考え方を認めるべき時期なのだろうと思う。

(ただし従来のテスト理論が間違っているとか時代遅れとか主張するつもりはまったくない。テストは所詮道具である。そして道具は適切に使われた時最大の効果を発揮するのである)。そのことによってのみ、様々な用途にテストが使用可能になり、社会的要請にも答えることができるようになるだろう。

引用文献

- [教心] = 日本教育心理学会第37会総会発表論文集
 [日心] = 日本心理学会第59会大会発表論文集
 [行動] = 日本行動計量学会第23回大会発表論文抄録集
- 麻生英樹(企画) 1995 シンポジウム「ニューラルネットワークとデータ解析」[行動] pp.63-71.
- Brehmer, B. 1979 Preliminaries to a psychology of inference. *Scandinavian Journal of Psychology*, 20, 193-210.
- Dawes, R.M., & Corrigan, B. 1974 Linear models in decision making. *Psychological Bulletin*, 81, 95-106.
- Falmagne, J.C. 1993 Stochastic learning paths in a knowledge structure. *Journal of Mathematical Psychology*, 37, 489-512.
- Falmagne, J.C., Koppen, M., Villano, M., Doignon, J.P., & Johanneson, L. 1990 Introduction to knowledge spaces: How to build, test, and search them. *Psychological Review*, 97, 201-224.
- Frederiksen, N., Mislevy, R.J., & Bejar, I.I. 1993 Test theory for a new generation of tests. Lawren-

- ce Erlbaum Associates.
- 藤井義久 1994 数学不安尺度 (MARS) に関する研究
教育心理学研究, **42**, 448-454.
- 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.
- 藤森 進・中野和代 1994 テストの教示が被検者の解
答行動と成績に与える影響 教育心理学研究, **42**,
454-462.
- Green, D.R., Ford, M.P., & Flamer, G.B. 1971 Mea-
surement and Piaget. McGraw-Hill.
- 萩生田伸子 1995 不安尺度の構成とその妥当性の検証
心理学研究, **66**, 16-23.
- 萩生田伸子・繁樹算男 1995 因子数選択のための新基
準の提案 [日心] p.439.
- 濱口佳和 1994 児童用主張性尺度の構成 教育心理
学研究 **42**, 463-470.
- 長谷川博一 1995a 自己評価の水準と項目評価性が再
検査効果に及ぼす影響 — 継時的比較理論の視点から
— 心理学研究, **66**, 141-145.
- 長谷川博一 1995b 項目再認率と再検査効果 [教心]
p.238.
- 平井洋子・渡部 洋 1994 小論文評点のカテゴリ化に
関する測定論的考察 行動計量学, **21**, 21-31.
- 池田 央 1994 現代テスト理論 朝倉書店
- 井上俊哉・孫 媛 1995 日本における差異項目機能
(DIF) 研究の可能性(2) [教心] p.233.
- 柏木繁男・山田耕嗣 1995 性格特性 5 因子モデル
(FFM) による内田クレペリンテストの評価について
心理学研究, **66**, 24-32.
- 狩野 裕(企画) 1995 シンポジウム「共分散構造分
析の基礎理論」[行動] pp.27-39.
- 北尾倫彦(編) 1995 中学校 思考力・判断力 その考
え方と指導と評価 図書文化社
- 木内亜紀 1995 独立・相互依存的自己理解尺度の作成
および信頼性・妥当性の検討 心理学研究, **66**,
100-106.
- Mayekawa, S. 1994 Equivalent path models in lin-
ear structural equation models. *Behaviormetrika*,
21, 79-96.
- 前川真一 1995 線形構造方程式モデルにおける同値モ
デル [行動] pp.34-37.
- Mislevy, R.J. 1993 Foundations of a new test the-
ory. In N. Frederiksen, R.J. Mislevy, & I.I. Bejar
(Eds.), *Test theory for a new generation of tests*.
Lawrence Erlbaum Associates.
- 武藤真介 1995 統計解析ハンドブック 朝倉書店
- 中野敬子 1995 女性を対象とした Type A 行動パター
ン測定法 — 日常行動質問表の作成 — 心理学研究,
66, 121-126.
- 大谷佳子・桜井茂男 1995 大学生における完全主義と
抑うつ傾向および絶望感との関係 心理学研究, **66**,
41-47.
- Ogasawara, H. 1995 Structural model of ability
distribution in the item response theory. *Behavio-
metrika*, **22**, 37-48.
- 苧坂真理子・苧坂直行 1994 読みとワーキングメモリ
容量 — 日本語版リーディングテストによる測定 —
心理学研究, **65**, 339-345.
- Palmer, S.E. 1978 Fundamental aspects of
cognitive representation, In E. Rosch, & B.B. Lloyd
(Eds.), *Cognition and categorization*. Lawrence
Erlbaum Associates.
- Pylyshyn, Z.W. 1984 Computation and cognition
— Toward a foundation for cognitive science. MIT
press.
(佐伯 胖(監訳) 1988 認知科学の計算理論 産業
図書)
- 繁樹算男・萩生田伸子 1995 選抜母集団における相関
係数のベイズ推定 [教心] p.243.
- 清水誠一 1995 絶対評価における客観化の試み
— ファジィ理論を適用した書写の評価法 — 日本教育
工学雑誌, **18**, 71-78.
- 下山晴彦 1995 男子大学生の無気力の研究 教育心理
学研究, **43**, 145-155.
- Siegler, R.S. 1976 Three aspects of cognitive devel-
opment, *Cognitive Psychology*, **8**, 481-520.
- Siegler, R.S. 1981 Developmental sequences within
and between concepts. *Monograph of the Society
for Research in Child Development*, **46** (2, Serial.
No.189).
- 孫 媛・井上俊哉 1995 日本における差異項目機能
(DIF) 研究の可能性(1) [教心] p.232.
- 菅原ますみ・島 悟・戸田まり・佐藤達哉・北村俊則 1994
乳幼児期にみられる行動特徴 — 日本語 RITQ およ
び TTS の検討 — 教育心理学研究, **42**, 315-323.
- 田栗正章・橋本明浩 1995 入学試験での合否入れ替わ
りに対する 1 考察 [行動] pp.46-47.
- 平 直樹 物語作成課題に基づく作文能力評価の分析
教育心理学研究, **43**, 134-144.
- 平 直樹・小野 博・前川真一・林部英雄・米山千佳子
1995 高校生程度の日本語能力テストの開発 — 語彙理
解テスト・漢字読み取りテストの尺度化 — 教育心理

- 学研究, **43**, 68-73.
- 高根芳雄 1995 制約つき主成分分析法—新しい多変量データ解析法— 朝倉書店
- Toyoda, H. 1994 A new identification rule and an estimator for the simultaneous equation model using the notation of the reticular action model. *Behaviormetrika*, **21**, 163-175.
- 植野真臣・大西 仁・繁樹算男 1994 教師の領域構造知識を利用した適応型テストの開発 日本教育工学雑誌, **18**, 15-24.
- VanLehn, K. 1990 Minds bugs : The origins of procedural misconception. MIT press.
- 矢口高士 1995 音楽作品の感情価測定尺度の作成および多面的感情状態尺度との関連の検討, 心理学研究, **65**, 463-470.
- 山下 元(編) 1995 ファジィ情報分析 —人間科学へのアプローチ— 共立出版
- 柳井晴夫 1994 多変量データ解析法 —理論と応用— 朝倉書店
- 吉崎静夫 1991 教師の意思決定と授業研究 ぎょうせい