

測定・評価部門

テストの理論と現実の「はざま」で

村 上 隆

(名古屋大学)

本稿では以下の3つの領域を主としてとりあげる。すなわち、第1に心理測定モデルとして注目されている項目反応理論 (item response theory, 以下 IRT と略記する。項目特性理論 item characteristic theory, 潜在特性理論 latent trait theory という名称も用いられる), 第2に入学試験, 第3に統計教育及び心理測定および統計の方法に関する啓蒙的書物である。

項目反応理論をとりあげるのは、日本教育心理学会第31回総会における、自主シンポジウム「項目特性理論の展開」が、きっかけである。これには私自身も指定討論者として参加させていただいたが、IRT という1つのモデルだけを話題にするシンポジウムに、これだけの話題提供者が集められることに、少々驚きを禁じえなかった。これ以外にも、何件かの個別発表があり、IRT は、いつの間にか日本の心理測定研究の中心的テーマとなった観がある。また、測定モデルを扱ったシンポジウムの開催も久々のことである。これはどうしても本稿でとりあげざるを得ない。他方、入学試験を本学会の最近の研究動向と考えることは後述するようにいささか無理があるが、これは、大学入試センターによる共通一次試験の「得点調整」という「事件」がきっかけとなった。私自身は野次馬的に事態の推移を見ていたにすぎないが、心理測定の方法にかかわることがが、社会的に大きな問題になったことは、従来あまり例がないと思われるので、その記録の意味でも扱いたい。

実は、近年の私自身の関心が、心理測定の方法と社会・文化とのかかわり、といったところにあり、どうしてもそうした観点から問題をとらえる傾向にある。その場合、私が頭に浮かべているかかわり方は2通りある。1つは心理測定の方法で現代の教育問題にどうアプローチするかということであり、これはまあ研究者として当然の姿勢と言えよう。もう1つは、心理測定の考え方をどのような形で現在の社会の中に浸透させていくかということであるが、それには、いわゆるパラダイム転換を通じて、心理測定の「考え方」そのものが変わっていくという可能性も含む。現在の私は、この第2の「かかわり方」の方にむしろ大きなウェイトを置いて考えている。

ただ、このことは単に私の個人的関心のありかたというだけではなく、固有の対象領域を持たない当部門のやや特殊な性格にもよるものである。測定や評価の方法だけが閉じた系として存在することが原理的にありえない以上、他の分野との交流とともにいわゆる啓蒙活動は常にこの部門の課題としてあり続けてきたのである。

共通一次試験に関して言えば、もっと良い「調整」の方法は存在しなかったのかと考えるのが、まともな第1の「かかわり方」であり、そこでは測定の理論が問題になる。一方、調整が原因となってひきおこされた議論の内容や、ある調整の方法が社会的に受容されるか否かに関心をもつのが第2のかかわり方であり、そこでは、広い意味での統計教育や、心理測定の思想的、文化的背景といったものが表面に現れてくることになる。それが、『統計的「ものの見方」の普及に向けて』という第3節を設けた理由である。

もっとも、それには付帯的事情として、発表形式の変更により、総会中に発表者とのコミュニケーションが行いにくくなったこともある。また1ページに短縮された論文集の内容を、単に要約することの意味はあまりあるまい。ただ、私のフィルターを通らない発表をすべて無視するのは、あまりに偏りが大きすぎることになるだろう。したがって、それらについても紙面の許す限りとりあげることにした。ただ、その選択は私の独断によっており、また、総会当日は本稿に取り上げるつもりでかなり詳しくお話をうかがっておきながら、結局とりあげなかったものがかなりある。貴重な時間を割いていただいた方々にお詫び申し上げたい。

なお、以下においてとりあげる研究のうち、教育心理学会第31回総会の発表論文については〔 〕内に論文集の掲載ページを示し、末尾の文献には掲げていない。

1. 項目反応理論をめぐって

前記の自主シンポジウム「項目特性理論の展開」から始めよう。企画者は渡部洋 (東京大学) であり、以下に述べるような6名の話題提供者による5つの講演が行われた。指定討論者は、渡辺直登、前川真一と私の3名で

あった。

1.1 項目反応理論とは

前述のように、わが国の項目反応理論の研究者の層は厚く、今回の企画は時宜にかなったものである、というよりむしろやや遅きに失したと評すべきかもしれない。そうは言っても、本学会の大半の会員にとって、まだこのモデルは馴染みが薄いものであらうと思われるから、話題提供者の講演に先立って、野口裕之（名古屋大）による理論の概要の解説が置かれたのは、適切な処置であったと言える。ここでもまず、その解説を参考にしながら、IRTの概要を述べることにする。この理論についてもっと詳しく知りたい方には、芝・渡部（1987）の付録A（pp. 153～178）、Lord（1980）の part I（pp. 1～80）等をお勧めしておこう。

IRTでは、1つのテストで測定される個人の能力が1次元の潜在特性尺度で表現できると仮定する（1次元性の仮定）。テストの個々の項目に対する個人の反応は、正答（1）または誤答（0）として2値的にコード化されるが、各項目への正答確率は、潜在特性の非線型な単調増加関数である項目特性曲線で表現できると想定される。項目特性曲線の形は、多くの場合2つのパラメータをもつロジスティック曲線であると仮定される。2つのパラメータの一方は項目の困難度、もう一方は項目の識別力をあらわす（まぐれあたりの確率を示す第3のパラメータが追加されることもある。それぞれ2パラメータ・モデル、3パラメータ・モデルと呼ばれる）。また、パラメータの推定のためには、もう一つ局所独立の仮定というものがことになる。これは潜在特性尺度上で同じ位置を占める（つまり能力が等しい）被験者が、2つの項目とともに正答する確率は、項目特性曲線で与えられる個々の項目の正答確率の積で与えられる、という仮定である。

ここまでの段階で、あえて古典的テスト理論との類比を求めれば、潜在特性尺度は、各被験者が正答した項目数として定義されるテスト得点に、項目特性の困難度パラメータは項目の正答率に、識別力パラメータは項目得点－テスト得点間相関係数に、だいたい対応する。また局所独立の仮定は、項目得点を真の得点と誤差に分解するモデルを想定したとき、異なる項目の誤差が相互に独立であるという仮定に相当する。

古典的テスト理論の場合、正答確率のテスト得点への回帰の形態としては直線を想定していることになるが、これは従属変数である正答確率の値の範囲が0～1の間に制限されている以上、厳密には成り立ちえない。また、テスト得点の α 係数が、信頼性係数と一致するためには、いわゆる「本質的に π 等価」（あるいは弱平行測定、池田、1973）、という仮定が必要であるが、これはすべての

項目の回帰直線の傾きが一定であるという、さらに非現実的な仮定である。IRTでは回帰モデルを非線型のロジスティック関数に置き換えることによって、理論の展開がはるかに筋の通ったものになるばかりでなく、いくつかの有用な特徴がもたらされる。たとえば、

(1) 各項目のパラメータが受験した被験者の集団とは独立に定義される（項目得点－テスト得点間相関は、レンジ効果により、受験者集団の特性によって変動する）。

(2) 各被験者の潜在特性尺度値は、出題された項目とは独立に定義される（正答項目数によるテスト得点は、もちろん出題された項目の数や困難度によって変動する）。

(3) 測定の精度は、潜在特性の値ごとに（つまり被験者ごとに）テスト情報量として得られる（古典的テスト理論における信頼性係数は、すべての得点範囲において同一であると仮定されており、もちろん1つのテストに対しては1つの値が得られるにすぎない）。

といったことである。

こうして、IRTは、古典的テスト理論における、「本質的に π 等価」という制約条件を緩和した一般化であると見ることができ、また他方では、ガットマン・スケールに確率的ゆらぎを導入した一般化としてとらえることもできる（McDonald, 1985）。しかし、被験者のスコアを、正答項目数ではなく、潜在特性尺度上の1点として、確率モデルを通じて推定することにより、そのどちらをも越えた不変性の高い結果を得ることができるようである。もちろん、かなり面倒な計算が必要になるが、そのコストは上記(1)～(3)のご利益を額面通りに受け取るならば、それによって十分につくられるであらう。

1.2 自主シンポジウム「項目特性理論の展開」

シンポジウムのと個々の講演について簡単に紹介しておきたい。

大塚〔S21〕は、放送大学の単位認定のための試験問題のアイテム・バンク・システムの構築と、そのデータベースの中にIRTによる項目パラメータを含めておくことの効果について論じた。アイテム・バンク作成は、試験問題の極端な偏りをなくするとともに、各学期の合格率を一定の水準（50～60％）に揃えること、問題項目の選択、出題原稿の作成をシステム化、省力化すること、困難度、識別力の分析を通じて、実施結果を出題者にフィード・バックするとともに、適切な問題の再利用を可能にすること、という目的でなされた。その結果、難問、奇問の減少、合格率の安定、出題作業の合理化、結果の問題作成への反映、といった成果があったという。特に、IRTの導入は、事前に合格率の予想が可能になること、合格ライン付近での精度がテスト情報量により明らかに

なること、(後述の等化の手法により)学期間の困難度の差を小さくすることができたこと、といったメリットをもたらした。ただし、プリテストの実施が困難であること、1学期の項目数に制限があることにより、項目パラメータに関する情報が集めにくいこと、分析に必要な四分位相関係数の推定が困難な場合があること、項目パラメータの変動が少なくない項目があること、といった問題点も指摘された。

これは恐らく、制度的に実施されている大規模な試験に、IRTが適用され、運用された、わが国で最初のケースであり、それが一定の成功をおさめただけでなく、実際の運用上の問題点が具体的に示された点も併せて、大変貴重な成果である。これはまた、未来の試験のありかたを予示するものであるとも言えよう。ただ、あらゆる試験がこのように運用できるかどうかは、また別問題であるとの印象ももった。たとえば、問題の再利用を通じて、テストを等化するテクニックなどを、大学入試センター試験にとり入れることは、仮に問題を非公開にする等の処置(日本の受験事情からして、現状では到底維持できるはずがないが)をとったとしても、大きな困難があろう。しかし、そういった問題が云々されるほどに、これは具体的、実地的な研究であると評価できる。今後の成果が待たれるとともに、他の試験におけるIRTの積極的な利用報告が続出することが期待される。

藤森〔S31〕は、IRTを学校教育の場で活用するための環境整備に関するものである。実践的適用を可能にする条件として、読みやすいテキストの作成、項目プールの作成とそのデータベース化、パソコン上で運用可能なパッケージ・プログラムの作成、といったことがあげられ、講演者自身の作成した、PC 9801上で使用できる、3パラメータ・モデルのプログラムが紹介された。

実際のデータなしには測定理論の開発も改良もありえないし、意味のあるデータの獲得のためには、現場の協力が不可欠である、というのが講演者の基本的モチベーションのようである。これはこの種のモデルの研究における1つの重要なポイントの指摘である。そのためには、理論から生み出されるものの有用性が現場に実感されることが先決であろう。そこで、まず現場のレベルで使用可能なプログラムを提供することは、確かに良い戦略である。ただ、S-P表の普及過程を見ても、そこには単なるhow toだけでない、説得力のある一種の「思想」が伴っていたと思われる。現在のIRT研究には、その点がまだ欠けているようにも思われるのだが。

服部〔S31〕は、Raschモデルを用いた語彙能力の適応型テスト(adaptive testing)のための、項目プールの作成の話題である。Raschモデルとは、項目の識別力

をすべての項目について一定であると仮定する、いわゆる1パラメータ・モデルである。通常の2、あるいは3パラメータ・モデルに比してさらに制約条件は強くなるが、正答項目数としてのテスト得点が被験者の潜在特性尺度値の十分統計量になるという特徴をもつ。適応型テストとは、被験者の能力に応じて、自動的にその被験者に対して適切な困難レベル(正答率0.5程度)の出題をおこなうことにより、短時間で精度の高い測定を目指そうとするテストの方式である。このようなテストの方式は、前述のIRTの(2)の特徴によって可能になるわけである。そのためにはもちろん、事前に項目の困難度が知られている必要がある。そのためのプリテストとして、相互に項目の重なりを持つ10個のテストをそれぞれ異なる被験者群に実施し、共通の尺度上で全項目の困難度を求める方法が述べられたほか、パソコンを用いた適応型テストの再検査信頼性、折半法による信頼性、固定型テストとの相関が検討された。その結果、適応型テストの信頼性はかなり高いものとみなされたが、固定型テストとの相関は0.703で、やや低いと評価された。

非常な力作であり、Raschモデルを実際場面に適用する野心的試みとしても評価されよう。しかし、この制約条件の強いモデルを利用するにあたって、やや、モデルの当てはまりに対する評価が甘い点は問題であろう。たとえば、四分相関行列の因子分析の結果から項目全体を1次元とみなして良いと判断しているが、この手続は2パラメータ・モデルに対する適合性の検証にはなっても、1パラメータ・モデルに対するそれとしては不十分である。そして例えば、固定型テストとの相関の低さは、モデルの制約のためか、適応型テストという方法のためか、明らかにされることが望ましい。また、指定討論者の前川から、adaptive testingの場面をベイズ流に扱うのではないかという指摘もあった。

孫・芝〔S31〕は、第2言語としての日本語学習者に、従来、日本人のために開発された語彙能力試験を実施すると、著しく逸脱した反応パターンが得られることを見出した。そこで、このようなデータによりふさわしいモデルとして、被験者の能力の変動を考慮にいたった一般的な項目反応モデル(GIRT)を新たに開発した。データへのあてはまりは、外国人のデータの場合、通常の2パラメータ・モデルに比較して著しく改善している。

理論的に大変野心的な試みであり、それを実際のアルゴリズムとして実現した努力は、大いに多とされよう。また実際、外国人のデータへの適合度の改善は劇的であり、この局面におけるこのモデルの適切性を物語っている。しかし、それにもかかわらず、外国人のデータの不適合を個人内の能力の変動に帰することには、私などは

強い抵抗を覚えざるを得なかった。第2言語の習得過程は、第1言語の場合とは全く異なるはずであり、学習者の語彙の習得の順序（それが大体困難度に相当すると思われる）は、母国語話者と異なって当然だからである。その違いは、個人内変動といった「可変的」要因に帰されるべきものではなく、むしろ全く異なる潜在特性次元として表現されるべきものなのではあるまいか。それとも、私のこのような見方は、モデルを実体化しすぎる「誤り」を犯していることになるのであろうか。また、シンポジウムの中で前川が指摘したように、GIRTでは、IRTに含まれるパラメータは所与とし、第2段階で、被験者の能力の変動のパラメータを推定しているが、これらを全体として同時に推定できないか、という問題もある。講演者の回答は肯定的であった。是非その方向での展開も期待したい。

野口〔S31〕は、IRTによる尺度の等化(equating)の問題を包括的にとりあげた。一般にテストの等化とは、「2つのテストの得点を同等かつ交換可能にするような変換システムを定める過程である」と定義される。この問題は、古典的テスト理論の範囲内においても様々な方法が提案されていた。しかし、前記のIRTの特徴、(1)と(2)をともに生かすことにより、尺度の等化を行うことができる。尺度の等化とは、個別に構成されたIRTの尺度を、共通の尺度に変換することであり、これによって、広い範囲の発達段階にわたる個人差が1本の物差で評価できるようにしたり、異なった時期に実施されたテストの得点を相互に比較可能なものにすることができる。野口の講演は、基本的な定式化、等化のためのデータ収集のデザイン（共通項目デザインと共通被験者デザイン）、等化係数の推定方法、結果の比較等について、自身の研究も含め、広汎にレビューすることに重点を置いたものであった。

最新の情報まで含めて、さまざまな方法が（時間の制約の中で）ていねいに紹介されていた。また、複数の方法を比較する際の、いろいろな視点が具体的に示されたことは有益であった。

シンポジウム全体を通して見ると、バランスのとれたテーマ選択と適切な講演者の選任がなされていたと感じられる。前述のように、この種のシンポジウムは、専門分野を同じくするオーディエンスを想定したオリジナルな研究報告と、それをめぐる討論だけに限定することは難しく、必ずしも測定・評価を専門としない会員に対する、一種の「啓蒙活動」という機能も併せ持つ必要がある。今回は、そのバランスもよくとれていたと思われる。また適当な時期を見計らって、同様な企画がなされることを期待したい。

1.3 項目反応理論に関するその他の研究

まず、一般発表の中からIRT関連のものを紹介しておこう。

喜岡〔392〕は、小学校の教育課程における整数、小数、分数の四則演算の能力の測定のための共通尺度をIRTを用いて構成する試みである。プログラムLOGIST5を用いて、2パラメータ・モデルによる項目パラメータの推定と等化が同時に行われている。モデルの前提条件、適合度についてもていねいに検討されている。従来、IRTの適用領域としては語彙能力が選ばれることが多かった。語彙は比較的構造が緩やかであり、IRTモデルに乗りやすい特徴をもっていたと思われるが、計算能力は喜岡自身も指摘しているように階層的構造的な強い領域である。このため、かなりモデルに対する不適合を生ずるのではないかと予想されるが、むしろ結果は2パラメータ・モデルがよく適合した。それでは、階層性はどのように表現されることになるのか。今後の検討に期待したい。

服部〔398〕は、前記、服部〔S31〕の中の項目プールの作成の部分をやや詳細に論じたものである。中学生から大学生までの範囲を覆う共通尺度の作成には、やはり等化の手續を要する。ここでも10個のテストの等化が試みられた。3つ以上の尺度の等化には、野口〔S31〕も指摘するように、誤差の累積の問題がある。その問題を解決するために、10個のテストを共通項目のネットワークで結び、服部自身の開発した等化誤差平方和最小法を用いて等化を行うというのが、この報告のポイントである。

繁樹・市川〔400〕は、CAIにおける教授系列の最適化を目指して、IRTモデルを種々の形で一般化する試みである。たとえば、被験者がMaster, Non-master, Partial Knowledgeの3群からなると想定し、当該領域をすでに習得済みのMaster群は、全項目に確率 $1-\delta$ で正答し、全く習得していないNon-master群は全項目に確率 δ で正答し、部分的知識をもつ第3群は、通常の2パラメータ・モデルにしたがって反応すると考える。これは、IRTに潜在構造分析の特徴を組み込んだという点でも、また、「習得」とは何かということを改めて考えさせる点でも、実に興味深い拡張である。このモデルが適切に当てはまるデータに関しては、たとえば、criterion referenced testにおける合否判定の恣意性は、ある程度解消されることになる。ただ、実際の学習過程において、学習者が截然とこの3つの群に分割されるということの概念的正当性が、いまひとつ納得できないのも事実であろう。実データにおける検討が待たれる。

山崎〔401〕は、直接IRTを用いているわけではない

が、明らかにその方向が目指されているのでここで扱っておく。中学校の定期テストと実力テストから、統計的手続を経て選択された145の項目の難易度と識別力を、正答率と（四分位相関係数を用いた）因子負荷によって定めた。一見、共通被験者デザインによる等化の手続に見えるが、問題は、このデータが約1年間にわたって同一の被験者からとられたものであることである（定期テスト等を利用してデータ収集すればどうしてもこうなる）。この間に被験者の能力は当然変化しているはずだから、少なくとも正答率としての難易度は共通の尺度上にあるとは言えない。

次に、IRTに関する論文をとりあげておこう。

岩崎（1989）は、正答を1、誤答を0とコード化する通常のデータ行列に、これとは逆に正答を0、誤答を1とした行列をならべてできる「拡大データ行列」を林の数量化Ⅲ類で分析し、項目の正答に付される重みを x 、誤答に付される重みを y として、 $x-y$ を項目の弁別度、 $x+y$ を困難度と呼ぶことを提案している。その上で、これらと2パラメータモデルのそれぞれの対応するパラメータとの関係を明らかにしている。その意味で、この論文はIRTを正面からとりあげたものではない。実データを分析した例は示されておらず、たとえば極端に正答率が高い（あるいは低い）項目において y （あるいは x ）が極端な値になることが多いという数量化Ⅲ類の特性を考えたとき、先の弁別度、困難度の指標がうまく機能するかどうかという心配はある。しかし、数量化の方法とIRTの関係に関して理論的に検討した例は、私の知る限り過去に例がなく、重要な研究と言える。

渡辺（1989）は、IRTを組織行動の測定に適用する可能性とその問題点についてレビューしている。たとえば、適応型テストを企業内試験や組織開発プログラムに適用したり、項目バイアスの分析や質問項目の翻訳の妥当性の検討にIRTを用いる可能性が指摘されている。確かに、IRTには狭い意味での教育測定を越えた適用可能性があり、こうした議論は有意義である。ただし、渡辺自身も指摘しているように、IRTの有用な特徴は、実のところその強い仮定からもたらされているのであり、その点の十分な吟味を怠らないようにする必要がある。とくに、研究のtoolとしてならともかく、それが組織の意思決定や、個人の処遇に関する何らかの決定に用いられる場合には、一段と慎重に対処する必要がある。

1.4 項目反応理論への疑問

私自身はこのような点も含め、先の自主シンポジウムの指定討論者としてもIRTに対する懐疑的な見解を述べた。また、一般発表としても、村上〔393〕において、心理測定尺度の1次元性が、現実には成立しがたい条件

であることを述べて、基本仮定に対する疑義を提した。実のところ、IRTによってもたらされるメリットは、IRTのモデル自体によるよりも、むしろ1次元性の仮定による部分の方が大きいと私は思う（もちろん異論はあろう）。たとえば、IRTの意味で1次元性が成立すれば、尺度の信頼性の問題は妥当性の問題と等価になる。つまり、1次元であることが確認されている項目のプールがあったとき、そこから尺度の信頼性が高くなるように項目選択をしていけば、自動的にその尺度の妥当性も高まる。だからたとえば、McDonald（1985, pp. 214～222）は、（非線型因子分析の）因子を構成概念と同義とみなしたうえで、全項目の共通性の和を（正答数の合計としての）テスト得点の分散で割ったもの（ ω 係数）は、テスト得点の信頼性係数であると同時に、妥当性、一般化可能性（generalizability）の係数でもあると論ずるのである。しかし、うまく1次元性が成立した（帰無仮説が棄却されない）としても、「その」次元が測定したい次元であることを自動的に保証するわけではないことも当然である。ましてや、実際には多次元である項目選択には、やはり従来通りの困難はつきまとう。ともかく、IRTを使用すれば、妥当性の問題が一挙に解決するわけではない。

厳密な意味の1次元性が実データでは成立しないという事実それ自体を、多くの研究者は否定しないであろう。実際的に重要なのは、実用的な観点からみたモデルの許容度でありこれは一種の匙加減の問題である。しかし、モデルが美しいものであればあるほど、現実がモデルを通してしか見えなくなるという傾向が確かに存在するように思う（トーマス・クーンの意味におけるパラダイムが、かえってものを見えなくする効果である）。その結果として、モデルにうまく当てはまるような項目だけが選択され、実際には極めて狭い（すなわち普通の意味での妥当性の低い）尺度が作られてしまう可能性が高い。

もちろん、IRTが心理測定に対する新しいパースペクティブを開いたことを否定するつもりはない。たとえば、測定の対象となる能力の水準によって測定誤差の大きさが異なりうることは、IRTがもたらした極めて重要な認識である（これも当たり前の事実と言えばそれまでののだが、古典的テスト理論の枠組みに執着している限り見えにくい）。しかし他方において、IRTがテスト、人間の能力、さらには個人差一般についてのイメージを、著しく制約してしまっている面を無視すべきではないと思うのである。

2. 大学入試をめぐる

2.1 入試研究

このところ本学会の研究発表としては途絶えているが、日本行動計量学会が、「行動計量学」誌の第15巻1号(1987年9月刊)で「選抜試験をめぐる」という特集を組んでいる他、第16回大会(1988年)ではシンポジウムを実施、第17回大会でも特別セッションを設けている。これらの中では、まず、肥田野(1987)が選抜試験の改善の問題を論ずるにあたって、基本的に抑えるべきポイントを網羅的に扱っており、入試問題を論ずる人の必読文献と言える。宮沢(1987)は、物理学者としての観点からテスト得点の分析モデルを考察している。教育心理学のテスト理論とは、その前提や扱いが異なっている点に興味深い。清水(1988)は、共通1次試験の理科、社会の難易度の科目間差の回帰モデルによる推定を含んでいる。鈴木(1989)は、傾斜配点の影響を実データを用いたシミュレーションによって評価しようとするものである(個人の得点を生成するのに2パラメータ・モデルを用いている)。

その他、渡部・平・井上(1989)は、昨年の本学会総会で発表され、先の行動計量学会のシンポジウムにも含まれていたが、小論文評価の問題を綿密に検討したものである。その結果、総合評価の評定者間の一貫性は極めて低く、評定者内一貫性も高いとは言えないこと、分析的評価のための諸観点が少なくとも内容に関するものと言語力に関するものに2分できること、分析的評価の第1主成分が総合評価と同程度の信頼性をもつこと、等を見出している。分析結果の資料も十分に付されており、貴重な文献である。なお、同じ個人が異なる時期に異なるタイトルで執筆した小論文に対する評価の安定性、といった問題を今後とりあげてほしいと感じた。

2.2 共通1次試験の得点調整

1989年1月21、22日の両日実施された共通1次試験のうち、理科の2科目、すなわち物理と生物について、1月27日、大学入試センターは、共通1次試験の開始以来初めての「得点調整」を行うことを発表した。共通1次試験の難易度の基準としては、ほぼ平均点が 60 ± 5 点の範囲に収まることが要請されているという。しかし、今回の場合各科目の平均点は、物理53.47、化学73.75、生物44.31、地学71.31と、最高と最低の間に30点近い開きがあった。これは、同時に行われた社会の各科目の平均点が、それぞれ、倫理・社会63.61、日本史66.46、世界史63.03、地理64.49、現代社会65.37であったことと比較すると確かに目立つ。最終的には文相の決断にもとづき27日夕発表された調整公式は、物理について $48.8 + 0.512X$ 、生物について $47.2 + 0.528X$ というものであ

た。この結果、物理の平均点は76.17、生物のそれは70.59に上昇する。これに対して、得点が修正されなかった化学と地学の受験者、およびその父兄からの苦情が相次ぎ、マスコミ論調もこの「調整」に対しては批判的であった。

入試センターのとった処置に関して、現在の時点で無責任に論評することは慎みたいが、やはりこの「事件」は、教育心理学の測定・評価部門として無視しえないものを幾つか含んでいると考える。

この「調整」の問題に関してはまず、理論的な正解が存在しないことを確認しておく必要がある。実際、この事態は前述の等化の問題としては成立しない。等化が可能であるためには、少なくとも2つの尺度が測定している次元が同一でなければならない(さらに、厳密には2つの尺度の信頼性が一致していなければならない)が、異なる科目の試験が測定しているのは、当然異なる次元のはずだからである。しかし、ある種の不公平の存在はあきらかなのだから、ここでは、理論的により良く、かつ社会的に受容される調整方式が選択される必要があった。まさしく私が最初にあげた測定の方法と社会との2つの「かかわり方」がともに問題になる。

そこでいくつかの選択が存在する。まず、得点の変換方式として、素点の何らかの線型変換による方法をとるか、パーセンタイルで調整するかということがある。次に、すべての科目の平均値を一致させるか、他の科目の得点の情報を用いて、科目ごとに異なる平均値を設定するかということもある。たとえば、容易に思いつく科目ごとの偏差値化は、すべての科目の平均値を一致させる線型変換である。ところで、私の周囲にも、なぜこの偏差値化の方法が採用されなかったのかを、いぶかしむ声があった。しかし、発表された得点分布の教科間の著しい違いを見れば、これが不適切な処置であることは明らかである。特に平均値の高い化学の分布は大きな負の歪みを示しており、満点の偏差値は他の科目より著しく低くなったはずだから、もしこの方式が採用されていたらその「不公平」が、もっと大きな非難をまき起こしたであろう。パーセンタイルを用いれば分布型の違いの問題は解消するが、調整公式が簡単な形にならず、得点の対応表を発表しなければならないことになる。しかし、恐らくこれが採用されなかった理由は、調整後の得点が素点より低下する受験者が生ずることに対する社会的批判への配慮からではなかっただろうか。私のこの憶測が正しいとすれば、ここでは、方法としての適切さよりも社会的受容可能性の方が優先されたわけである。変換後の得点が変換前より低下する受験者を出さないためには、先に見たような、100点は100点に変換されるという「制約条件」付きの線型変換がほぼ唯一の方法である。

しかし、定員が厳守される現行制度の下では、誰かの得点が上昇すれば、他の人の得点はかならず相対的に低下することになる。多くの人々はこの単純な事実にも気がついていないようにみえる。したがって、最初は、物理、生物の受験者に対する「不公平」が言いたてられ、調整後は、得点が調整されなかった化学、地学の受験者に対する同情論が強調される、ということになった。入試センターの配慮はむしろ仇になったと言うべきか。

実際には、以下のような問題も存在するはずである。
 (1)あまり目立たないが、選択でない科目についても実際には不公平が存在する。たとえば、英語(110.07)と国語(139.01)の平均点の差は20点近い。このことは、国語の得意な受験生にとって不利なはずである(相対的な差を得意科目で大きくできないから)。さらに、分散の大きさとか、高得点部分での信頼性の低さ(IRTの傾分である)といったこともあげつらう余地がある。
 (2)調整の対象となった2科目のうち、少なくとも生物に関しては良問であったと評価されているようである。実際、単なる孤立した知識の記憶では、こなせない問題であり、平均点が低かったのは多くの受験者が、こういう問題に対する準備をしていなかったことを意味していると思われる(こうしたことを見ても、1つの科目の学力が1次元であらわせるなどという想定が、ほとんど虚妄に近いことがわかる)。それでは、こうした出題の「改良」を今回のような大幅な平均値の低下をまねくことなく行うにはどうしたらよいのだろうか。ここには、少なくとも、応用・分析の能力の具体的な記述方法という教育目標の分類学の問題が絡んでくるし、試験の内容の事前予告の形式も問題になる。現在のところ、試験のこの種の側面については、過去の問題から予測するしかないからである。

いずれにせよ現状では、こういった問題について議論できるような社会的土壌は、まったくできていないと考えざるをえない。前述の「行動計量学」の特集の最後に、池田央(立教大)による次のようなコメントが付されている。「問題は現在のところいろいろな選抜方式のもつ側面に対して無知であるというよりは、むしろそれぞれの制度がもつ功罪を国民のすべての人が情報として必ずしも共有していないことにある。……ここでとりあげられた識者の議論が、さらにより広い層への共有財産として普遍化していくことを願ってやまない。」ここには、まさに入試問題のみならず、測定・評価部門の大きな問題の指摘があるように思える。

3. 統計的な「ものの見方」の普及に向けて

「教育心理学研究」誌を少し眺めてみただけでも、測定

の技術、あるいはもう少し拡げて統計的技術が、教育心理学の研究にとって今日不可欠になっていることは見てとれる。一方、それらの背景にある考え方が、どの程度浸透しているか、ということになると少々疑問である。しかし、数式だらけの教科書から「意味」を汲み取るとは、「国民のすべて」にはもちろん、研究者にとっても決して容易なことではない。

こうした局面からの脱却をはかる1つの方向を示した書物として、まず市川(1988)をあげたい。大多数の人間が持っている統計的直観のバイアスをまず明らかにし、それを視覚的イメージを用いながら正しい方向に導こうという認知科学的方法の提唱である。市川(1989)は、これに実践的学習法と認知カウンセリングを加えて、現在の大学レベルの統計教育を高校以下に下ろしていこうという、さらに大胆な提案である。教育行政・制度にはまったく素人の私の目から見ても、この提案の実現が当面極めて困難なのははっきりしている。しかし前節でも指摘したように、現在の試験に対する一般の議論の水準の低さの原因の1つが、この種の統計的知識の普及の不十分性にもあることはあきらかであり、その一方で、そのための具体的提案が、教育心理学者によってなされてこなかったのも事実である。また、従来の統計学の知識の表現形態が、その普及を妨げていた可能性はおおいにある。市川の言う(正しい)統計的直観が、より充実し、かつ洗練されていくことを期待したい。

渡部(1988)は、久しぶりにあらわれた多変量解析のテキストである。従来のものと比べて読みやすく、教えるやすい。従来あまりなかった実際の適用上の問題(主成分分析と正準相関分析を併用した場合の結果の解釈等)にも触れられている。ただ、数式がイタリック体で表記されていないのは(私にはかなり抵抗があるが)何か理由があるのだろうか。

鷲尾・大橋(1989)は、品質管理を対象領域として想定しているために、心理学者にはややとりつきにくいのが、従来の手法別の解説とは一味違った、いわば課題指向型の入門書である。記述のスタイルは、先の市川(1988)の発想に近く、直観的、イメージ的表現がうまく取り入れられている。

池田(1989)は、公式集と事典の中間的形態の新しい形の書物と言えよう。内容的にも新しいところまでフォローされており、今後お世話になることが多いであろう。ただ、前書きに述べられている「統計値の背後にある本来の意味を正しく理解する」という目的でこの書物を活用できるのは、相当にレベルの高い読者なのではあるまいか。

吉村(1987)は、異分野(薬物と毒物の効果の検証)

の、かつ統計的検定に限定された書物だが、ぜひとりあげておきたい。たとえば、なぜ分散分析にはさまざまなモデルがあって「正しく」使いわけの必要があるのか、3つ以上の条件間の比較を、なぜ1検定の反復でおこなってはいけないのか、納得のいく説明をまだ聞いたことのない方は一度お読みになるべきである。このスタイルの「心理・教育統計学」が、近い将来出版されることを期待したい。

もう1つ、将来是非とりあげる必要がある分野に、統計ソフトウェア・パッケージとその解説書がある（レビュー期間の出版物としてとりあえず高橋・大橋・芳賀、1989をあけておく）。鷲尾・大橋（1989）も指摘するように、あまり品質の高くないパソコン用統計ソフトの氾濫による、統計解析の水準の低下は否定できない。できるだけ早い時期に、その種のソフトの「総点検」といった作業が必要であろうし、また、信頼度の高いパッケージについての信頼できる解説の必要性も高い。

次の2つは、より一般的な書物であるが、とくに知能の測定に関する議論が、心理測定理論の暗黙の前提を明確にしている点で、この部門にとって有益である。

グールド（1989）は、教育心理学の研究、特に因子分析の理論とその適用とを十分踏まえて書かれた、生物学者による知能論である。研究法の中に忍び込む生物学的決定論という偏見、因子を実体化してしまう（著者の言う具象化の）誤り、それから直線的進歩観が作り出した知能観に対する説得的な批判である。因子分析の方法の解説の仕方も、図的直観と巧みなアナロジーが用いられ、これも市川（1988）の主張と近いところが興味深い。

東（1989）は、知能について、さまざまな観点が「心理測定的研究」と対置されている。他の分野の示す知能への柔軟なアプローチと、東の示す「さまざまな文脈における知能」のイメージは、心理測定の立場からも学ぶべきものがある。

この2著に共通して見られる特徴は、相関的、多変量的アプローチを、積極的に記述に取り入れている点である。これにより、通常のテキストや通俗的読み物よりはるかに深い水準の議論が可能になっている。このあたりの議論の仕方（内容でなく）が、教養課程の心理学の常識になることが期待されよう。

4. その他の研究

豊田（1989）は、相関構造分析モデルを、Z変換された相関係数行列に対して、最小2乗法の意味であてはめるというものである。最急降下法のアルゴリズムが導出され、シミュレーションによって、通常の（相関行列に対する）最小2乗法（ULS）と推定精度が比較されてい

る。標本の大きさにかかわらず、一般に因子負荷の真値が大きい場合には、ここで提案された方法の精度が、通常の最小2乗法を上回ることが示されている。

モデルの扱いについての著者の力量は確かであるが、この方法のメリットは必ずしも理解しやしくない。ULSの問題点は、Jöreskog（1978）が述べるように、「標本共分散行列 S の要素である分散と共分散は、一般に相互に相関しており、分散も等質でないから、ULSは S と母共分散行列 Σ の間のズレを測るには間違ったメトリックを用いているのだ」という点にある。Z変換により、確かに分散は等質化されるが、より重大なデータの相互相関については問題が解決しないからである。ともかく、この方法では、モデルの当てはまりに関する検定方式を導くこと、パラメータ推定の標準誤差を算出すること、といった推測統計学的取り扱いはできない。この点は、通常の最尤法や一般化最小2乗法による共分散構造分析と比較して、明らかに劣ることになる。この方法が市民権を得るためには、たとえば、計算量が少ない割に、最尤法に近い精度が得られる、あるいは、小標本の場合にこの方法の方が何らかの点で優位にある、といったことが示される必要がある。

椎名・浅川〔396〕、浅川・椎名〔397〕は、先験的に定式化された濃度比較課題に対するストラテジー（ルール）から、課題に対する正誤パタンの確率モデルを導き、最尤法を用いて、個々の被験者の採用しているルールを推定しようとする方法論の提案と、それをういた実データの分析とからなっている。明らかに従来の仮説検定とは異なった研究のアプローチが示されている。現在のところ適用できる領域は限られていると思われるが、この未来指向的姿勢は賞賛に値しよう。

松田・浅井・大村・外島〔389～391〕、杉原他〔402, 413, 414〕の、前者は情報処理技術者の適性テストについて、後者は集団式認知スタイルテストについて、昨年から継続して信頼性と妥当性の検討を行っている。テストのこうした多面的で地道な研究の価値は決して過小評価されてはならない。

田中〔412〕は、学校教育の中でなされるソシオメトリック・テストの問題点を指摘し、改善策を提案している。選択の人数制限を行わないこと、排斥の記入は被験者の自由に任せること、処理のためのパソコンプログラムの開発、（人数制限の撤廃にともなう）新しい指標の考案等、どれもうなずけるものである。

5. おわりに

いささか「こころあまりてことばたらず」の記述となつたが、もはや紙面も時間も尽きた。あと一言。たまた

まのぞいた自主シンポジウム「発達研究において個人差をどうとらえるか？」の中の、氏家〔S32〕の議論を印象深く聴いた。私なりに要約すれば、(1)被験者の反応は、それ自体が単独で意味をもつわけではなく、その被験者の生活する文化的文脈の中で解釈されなければならない、(2)それを解釈しようとする研究者の側がもっている文化的前提を、研究者は常に意識化する努力を怠るべきではない、といったことになろうか。現在の測定・評価部門の研究に欠けているのは、このような文化的前提を認識しようとする姿勢であり、そのことは、理論の精緻化や実用性の追求と、少なくとも同程度に重要なのではないか、というのがどうやら私の言いたかったことなのである。

引用文献

- 東洋(柏木恵子編) 1989 教育の心理学—学習・発達・動機の視点 有斐閣
- ゴールド S.J. 鈴木善次・森脇靖子(訳) 1989 人間の誤測りまちがい—差別の科学史—河出書房新社 (Gould, S.J. 1981 *The mismeasure of man*. New York: W.W. Norton.)
- 肥田野直 1987 選抜試験に関する諸問題 行動計量学, 15—1, 24—33.
- 市川伸一 1988 決定における規範的推論と直観的推論 小橋康章 認知科学選書 18 決定を支援する 東京大学出版会.
- 市川伸一 1989 情報教育としての統計教育 日本教育工学会第5回大会予稿集, 113—116.
- 池田 央 1973 心理学研究法 8 テストⅡ 東京大学出版会.
- 池田 央(編) 1989 統計ガイドブック 新曜社
- 岩崎 学 1989 数値化Ⅲ類によるテスト得点の解析 行動計量学, 17—2, 13—21.
- Jöreskog, K.G. 1978 Structural analysis of covariance and correlational matrices. *Psychometrika*, 43, 443—477.
- Lord, F.M. 1980 *Applications of item response theory to practical testing problems*. Hillsdale NJ: Lawrence Erlbaum Associates.
- McDonald, R.P. 1985 *Factor analysis and related methods*. Hillsdale NJ: Lawrence Erlbaum Associates.
- 宮沢弘成 1987 入学試験の科学化にむけて 行動計量学, 15—1, 34—42.
- 芝 祐順・渡部洋 1987 入試データの解析 新曜社.
- 清水留三郎 1987 国立大学と公立大学による入学者選抜試験の改善のための統計的分析の事例 行動計量学, 15—1, 43—52.
- 鈴木規夫 1989 入試における選抜効果に関するシミュレーション研究 第17回日本行動計量学会大会発表論文抄録集, 3—4.
- 高橋行雄・大橋靖夫・芳賀敏郎 1989 SAS で学ぶ統計的データ解析 5 SAS による実験データの解析 東京大学出版会.
- 豊田秀樹 1989 相関構造分析におけるフィッシャーの Z 変換を利用したパラメータ推定法 教育心理学研究, 39, 67—70.
- 鷲尾泰俊・大橋靖雄 1989 入門統計的方法 3 多次元データの解析 岩波書店.
- 渡部 洋(編著) 1988 心理・教育のための多変量解析入門 福村出版.
- 渡辺直登 1989 項目反応理論を用いた組織行動の測定—その概要と適用可能性—経営行動科学, 4, 65—74.
- 吉村 功(編著) 1987 毒性・薬効データの統計解析—事例研究によるアプローチ—サイエンティスト社.