討論論文

テスト研究における理論と現実のバランス

名古屋大学教育学部　**野口裕之**

今回の誌上シンポジウムでは「テストの常識と非常識」というテーマの下に，3人の方々が提案者として論文を寄せられている。論文1では，統計的方法の基本ではあるが実際に国立大学の入試研究の場で十分に配慮されないことがある問題を鋭く指摘・解説された。論文2では，従来の教育心理学では簡便なモデルで済ませていた多枝選択形式項目に対する受験者の応答に対して，より精密なモデルを提案された。論文3は，教育心理測定理論をこれまでと全く異なる前提から出発し，構築する意図を持ってその基礎となる考え方を提案された。

3つの論文はそれぞれに極めて重要な指摘をしておられるが，読者にとってテストに関わる研究がいかに多面性を持つかが分かる。3名の提案者のうち2名の方が教育心理学以外の専門である事もテスト研究の広がりを示すものである。何れの論文も興味深く読む事ができたが，3つの論文をまとめて統一的に討論するのは，討論者の力量の所為もあって難しいので，各論文毎に取り上げることにする。

1　「国立大学の入試に関する常識と非常識」に対して

平野氏は，先ず「1　はじめに」で，テスト得点は実力の近似値であり，変動するものであるとし，「2　自己採点方式」で，共通一次試験の成績は実力そのものではないし，一般に複数回測定した結果の平均値の方が変動の幅が小さく，より正確に実力がわかるから，受験生が1回きりの共通一次試験の自己採点結果ではなく，いくつかの資料をもとに志望校の決定をしてくれると，“まぐれ”で合格する受験者が減るので，自己採点方式は止めるべきであると主張している。“テスト得点が変動するものである”ことを“常識”にしなければならないということである。このことは，教育心理学研究者の間では，比較的“常識”になっているが，古典的テスト理論のモデルを通した理解が十分に浸透しているとまでは言い切れない。というのは，学部レベルは言うまでもなく，大学院レベルでも古典的テスト理論を含む教育心理測定理論の教育が満足に行なわれることが少ないという現実があるからである。「3　受験機会の複数化」では，測定精度を上げる為には，複数回のテストそれぞれの上位を選ぶ方法もあり，そのためには受験機会の複数化は望ましいが，大学側の負担増は望ましくないし，その大学に望ましい能力を試験で測定するのだから，複数回の試験で別の能力を測定するというのも特別な場合を除いておかしいとしている。今回のシンポジウムのテーマとは直接関係ないかもしれないが，大学側の“負担”ということも入学試験を考える上で大切な要素である。入学試験は大学全体が莫大なエネルギーを使って実施するものであり，それを複数回繰り返すというのは関係者に大変な負担を強いることになる。確かに入学試験は大学にとって重要な教育活動の一つではあるが，大学はそれだけをやっている訳にはいかない。受験生にとって受験機会の複数化は望ましいことであるが，大学側はその業務全体を見渡して最適解を考えなければならない。このことは“常識”であり，誰もが理解していることであるが，現実には，“受験生のため”という“たてまえ”におされて複数回入試を実施している大学が多い。しかしながら，受験生あるいは学生に対するサービスはトータルに考えてより良いものにすべきであり，入学試験だけを取り上げても仕方がない。平野氏の主張するように受験機会の複数化は大学間で実施すべきものである。

「4　選抜効果」では，①ｘとｙの両方で選抜された者を対象とする場合　②ｘまたはｙの一方で選抜された者を対象とする場合　③ｘとｙの何れでもない，ｚで選抜された者を対象とする場合，の3つのタイプの選抜効果について，相関係数の解釈に対する注意を喚起している。この問題は，例えばGulliksen（1950）など古くから教育心理測定の専門家の間では知られていた問題であるが，通常の統計的方法の入門水準のテキストに記述されることはほとんどない。

「5　回帰現象Ⅰ」では，実力のテスト結果への回帰を取り上げ，テスト結果のトップグループは，たまたま実力を上回った者がほとんどを占め，実力的にトップグループを形成しているとは言えず，逆にボーダーライン上の者の中に，実力がトップに近い者が含まれている可能性が高いことを明示し，変動に特別な理由を付けて説明する誤りを「回帰の錯誤」と呼んで警鐘を鳴らしている。「6　回帰現象Ⅱ」では，高校側の常識である“センター試験では，浪人しても伸びはあまり期待できない”というのに対して，大学側で“ある医学部では1年浪人すると100点程上がる”という見解があるが，これは前年度にたまたま実力を下回った得点しかとれなかったものが，実力の上昇があるとしても，主として回帰の変動によって得点が上昇したものであることを示している。勿論，高校在学時にあまり勉強せず不合格になり，浪人中に勉強して実力を大幅に上昇させる例も少なくはないが受験生集団全体について考えるとその通りであろう。

ところで，ここで取り上げられた“非常識”は教育心理測定学の中で理論的には決して初等程度の水準を超えるテーマではない。それにもかかわらず，入学試験研究で“常識”となっていないのは，これらの問題をきちんと解説した入門書がほとんどないことに原因がある。例えば，池田（1976）では，選抜効果は相関係数の解釈上の注意の一つとして，回帰効果は前期試験と後期試験の間の変動を例として取り上げているが，東京大学教養学部統計学教室（1991）では両方とも記述が見られない。統計学一般に対する入門書では，むしろ記述されないのが普通である。そうだとするならば，これらの“常識”を身につけて欲しい教育心理学専攻の学生に対する統計教育ではテキストの選択を慎重にし，場合によっては補充教材を用意して対応しなければならない。ただ，問題は教育心理学専攻のカリキュラムの中で，教育心理測定学や教育統計学に十分なウエイトを置いているケースが少数派であることである。大学によっては半期1枠2単位分しか開講されない場合もある。そのような場合は，相関係数や回帰直線について一通りの解説を加えるのがせいぜいである。学生は基礎的な知識のほんの一部しか与えられないことになる。教育心理学研究者になるのはその中のほんの一部ではあるが，このような土壌からはなかなか“常識”が普及して来ない。

2　「ＭＣＱテストにおける受験者の Partial Knowledge の特性」に対して

有田氏は，「第1章　MCQテスト」で単一式，2連式など“MCQテスト”の種類を挙げている。教育心理学では通常，多枝選択形式テストと呼ばれることが多いが，単一式以外はあまり用いられることはない。特に，単一式変形で「e. ①～④のいずれでもない」という選択枝は，受験者が出題者が設定した正解よりも適切な正解を考えついた場合にこれを選択する可能性があるので，特別な理由が無い限り用いない方が良いとされている。しかしながら，有田氏の論文では，知っているか否かの2値的な状態がクリアな場合を想定しているので特に問題はない。

「第2章　2連式ＭＣＱテストの特性」及び「第3章　5肢択一式テストの特性」では，解答コードを構成する選択枝（有田論文の例では，5つの町村名）に対する知識の状態（知っている－知らない）の組合せを取り上げ，部分的な知識だけで正答できる知識パタンを“正答ターミナル”と呼び，また，当該選択枝に対する知識が正解の解答コードを同定するのに何ら寄与しない場合に“不要肢”，他の選択枝の知識を用いても難しすぎる為に当該選択枝にあて推量しないと正解が得られない場合に，“禁止肢”と呼び，“2連式”及び“5肢択一式”の場合について，実際にどのような場合がそうなるのかを解説している。筆者は，このような理論的考察の重要性を十分に認めるものであるが，テストを作成す

る場合には，“5肢択一式”での正答ターミナル（－ｘｘｘｘ）の存在とランダムゲシングによる正答確率が受験者の知識状態に応じて変化すること以外は，あまり問題にならない気がする。というのは，“2連式”に伴う問題は，選択枝の正誤パタンを無理に5つの規則性をもつ解答コードに押し込めることから生じている訳で，各選択枝毎に正誤を問うのがあて推量による正答確率が1／2になる事を嫌うのならば，正しい選択枝の番号を解答させるなど，出題形式を変えれば済む事と思われるからである。また，“5肢択一式”の禁止肢についても，テスト項目を作成する際には受験者の水準を考慮して“まよわし”を作る，すなわち犯し易い誤りを“まよわし”とする訳であるから，極めて難しい選択枝を入れるというのは，テスト項目の作成時から避けるように配慮すべき事と考えるからである。

「第4章　数学モデルによる正答ターミナル，不要肢，禁止肢」では，第2章，第3章で述べられた，正答ターミナル，不要肢，禁止肢について数学モデルを導入して理論的に導いている。

「第5章　m肢択一問題における知識の推移過程」では，確信を持って誤答する場合や選択枝に対する知識の状態があいまいな場合も含んだ，受験者の解答過程に関する包括的な数学モデルを示している。有田氏が最も力を注いだ箇所で，第4章までは第5章の為の準備とも言える。従来の教育心理学には全くない，いわば本格派のモデルで，筆者は興味深く読み進んだ。ただ，今回の論文では，このモデルを実際に適用する際に，知識の状態（知識の不完全さの度合い）や不完全知識からの推移確率（例えば，有田論文の図6参照）をどのようにして決めるかについては言及されていない。

企画者は，有田氏がこれまでにもこの内容について日本教育心理学会で発表されていにもかかわらず，教育心理学者からの反応がない事を指摘しておられる。確かにその通りである。専門分野の異なる研究者の発表に対して，当該分野の研究者からの反応が得られない例は“耶馬台国論争”にも見られるようである。有田論文の内容に関しては，一つには，“2連式”という教育心理学者にとって馴染みの薄い形式が取り上げられている事と測定・評価領域しかも数学モデルを用いる研究者は極めて少数であること　等が影響しているものと筆者は考える。

有田論文では，全体として精緻な数学モデルによりMCQテスト項目に対する受験者の解答過程を明らかにしようとしている。筆者は有田氏のモデルは，教育心理測定の分析モデルとしてよりもむしろテスト項目に対する受験者個人の解答過程を解明するための数理モデルとして有用であると考える。あいまいな知識状態や推移確率を変化させてシミュレートすれば面白い結果が得られよう。テスト研究の一つの在り方を示している。ただし，テスト得点の分布など，テスト全体や受験者集団について検討するには，知識状態パタンの分布を知る必要があるなど，さらにモデルを展開する必要がある。ただ，筆者はテストの作成，項目分析や得点の分析に用いる教育心理測定モデルとしては精緻に過ぎるように思われる。それはこのような研究の重要性を認めないのではなく，筆者にとって馴染みのある項目反応理論（項目応答理論）における最近のひとつの流れである“モデルの精緻化”と似ている気がするからである。項目反応理論では，あて推量を考慮に入れた3パラメタ・ロジスティックモデルがあるにもかかわらず，多枝選択形式項目から構成されるテストに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が用いられることがある。それは，テストを構成する項目の困難度は受験者の水準を考慮して決定される為，あて推量が大きく影響することは考え難いし，項目間で困難度が適当な幅を持つように構成される為，多くの項目であて推量による解答をする受験者はほとんど存在しない，等の実際的な判断があるからである。これに対して理論的な研究者は，さらに認知過程を組み込むなどパラメタの数を増やしてより複雑な項目反応モデルの開発を行なっている。モデルの精緻化である。その事自体は研究の発展の一つの方向を示しており，筆者も評価するものであるが，実際のテスト場面に適用しようとするとパラメタ値の推定というところで問題が生じてしまう。パラメタ数の多い精緻なモデル程，そのパラメタを推定するのに多数の実際の被験者の項目反応が必要になる。もし，少数の被験者数でパラメタ推定を実施したならば，その推定精度は保証されなくなる。モデルがより真実に近くなる程，パラメタ推定値の推定精度が怪しくなるというのである。実用的には，その辺りに適当なバランス感覚が必要である。

しかし，一方で精緻なモデルも様々なシミュレーション等によって有益な知見を与えてくれる。結局，モデルにどの程度まで現実の世界を忠実に反映する事を要求するかについては，モデルを適用する状況を踏まえて十分に検討する必要がある。筆者は有田論文の内容についても理論モデルとして評価し，筆者のように実際的な立場に立つ者の暴走を押さえる役割を期待するものである。

3　「家族的類似の概念にもとづくテスト得点のモデル」に対して

村上氏は，「1　個人差測定の問題」で通常実施されるテスト作成手順を復習し，その中で特にα係数を取り上げ，少数の等質な尺度を作成する事の問題点を指摘した。ただ，少数の次元で認識しようとするのは，ヒトの認識の仕方として普通の事である。“固定した”のが問題であるが，これは探索的な分析の場合には，“一時的に固定してみると”という程度のことで悪影響は言い過ぎの気がする。

「2　家族的類似モデル」で“特徴”という概念を導入し，“相互にほぼ無相関な複数の特徴を，項目間で共有しあうことによって項目間相関が生ずるという想定に依って立つモデル”を“家族的類似モデル”と名付け，これと従来の1因子構造のモデルとが同一の相関行列を説明し得ることを示した。「多数の特徴のうちの一つだけが，複数の項目によって共有されるといった仮定は，認めがたいものである。」と言うが，これは，データを巨視的にみるか微視的に見るかによるもので，認め難い仮定という程ではないであろう。また，「1因子，あるいは有限のある因子数が十分な標本の大きさの下で受入れ可能であるようなデータが存在したとしても，家族的類似モデルを対立仮説として考える限り，必ずしも，真の得点＋ランダム誤差，という形のモデルが立証されたことにはならないことを示すものである。」と述べているが，これは，モデルの立証と考えるからややこしくなる。（2－13）の場合もx_1とx_2，x_2とx_3，x_3とx_1の間で共通成分が含まれていることは確かである。一因子かどうかは別にして，x_1 x_2 x_3が相互に似ていることは確かである。だから，一つの“家族”として認識できる。一つに括れることの別の表現を（2－18）がしている。一因子モデルでは，唯一の因子が“真に”存在する事を立証しようとしているのではなく，変量群が一つに括れる事を示している。データ解析は真の構造を検証するものではなく，多数の変量間の関係を分類・整理する為の手段にすぎず，例えば，一つのグループに括られた変量群が真にどのような構造（ここで言う，一因子か家族的か）かは内容的・実質的な観点から判断すれば良い。理学的ではなく工学的な発想が教育心理測定の領域には必要と考える。

「3　家族的類似モデルの含意」では，さらに家族的類似モデルが“テストの作成や得点の解釈に対する示唆には無視できないものがある”とし，特に“領域範囲”という概念を導入し，内的整合性とは異なり，当該尺度が含む“特徴”の数，言い換えると尺度の“ふくらみ”を尺度作成にあたって考慮すべきとする。この中で，「領域範囲自体を明示的に表現するモデルが存在しなかった」と述べているが，これは尺度に含まれる項目群を分類整理し，純粋化することを目的としたモデルを作って来たし，実際要求されて来たので，逆に“ふくらみ”を持たせるようなモデルがなかったのであろう。また，「探索的因子分析の乱用を戒め，……確認的因子分析への以降を促進しようとする議論がある」と述べているが，いわゆるサイコメトリックスの研究者はどうしても統計的に洗練された方法の方に引きずり込まれるし，研究もそちらに偏りがちである。プログラムパッケージもすぐに普及し，領域研究者はその結果をつい鵜呑みにしてしまう。村上氏も言う通り，教育心理測定にとって大切なのはむしろ，探索的に，構造に様々な可能性を考えつつ実際に役に立つ尺度を作ることである。推測のプロセスだけが洗練されても意味はない。拠って立つ前提を満たすデータが入手できなければ話にならない。

「4　尺度の妥当性検討への含意」では，理論とデータとは相互に影響を与え合い変化していくものであるから，尺度の領域範囲も広くなったり狭くなったりするし，構成概念も変化し，従って，妥当性の検討は常に研究途上にあり，終わりが無いとする。

村上論文には，大きくは二つの意義がある。第一には，従来の古典的テスト理論や項目反応理論で重視される尺度の内的整合性に対して，いま一つ領域範囲という尺度のふくらみの概念を導入したこと，

そして，第二には，家族的類似モデルを導入することによって，現在の教育心理学研究者が通常行なっている，データ解析そしてその結果の解釈過程一般に対して警鐘を鳴らしていること，である。前者については村上氏も述べているように，“実践的には従来も強調されてきた”が，きちんとしたモデルをたてて説明されることは無かった。“領域範囲”を定式化する今後の研究の発展を期待したい。後者については，記述的方法よりも推測的方法が，例えば，探索的因子分析よりも確認的因子分析が，さらに共分散構造分析がより高級な分析法であり，教育心理学の研究により役に立つ方法であるという錯覚から目を覚まさせてくれる，という教育的な意義を特に強調したい。というのは，例えば，同一の相関行列に対して，一因子構造もあてはまるし，家族的類似構造もあてはまる事を示しているが，この事は，どんな仮説的な構造についても，仮説検定の結果棄却されなかったとしても，それが“真の構造”である事が示された訳ではないことを明らかにしてくれるからである。この事は，これまで仮説検定を教える際にも，検定の結果，“仮説が採択された”ということは積極的に“仮説が正しいことが立証された”事を意味するのではない事が強調されていたが，単に同一の構造で複数の値があてはまりうるというのではなく，複数の構造があてはまり得ることを示した点で画期的である。しかも，村上氏は，“真の構造”が検証されて研究が終了するのではなく，その“真の構造”というのがあくまで暫定的なものであることを強調する。これも，従来から言われてきた事柄ではあるが，家族的類似モデルの文脈の中に位置付けて主張するだけにより身近な問題として感じられる。見かけ上精緻な統計理論を用いても，決して“真の構造”に1回きりでは迫れない事を意識の上に上らせてくれるし，研究者を謙虚にしてくれる。

ただし，教育心理統計法の授業で最初からこのような話をすると，学生は悲観的そして懐疑的になってしまう。入門段階では，従来通り分析法をキチンと教え，ほんのさわり程度このような問題を教えるのが適当と考える。しかしながら，次の段階ではこのような事を十分に教える必要がある。たった1回のデータ収集・統計的分析で“真の構造”が確認できるものではなく，統計的方法もある一つの側面から切りこんだものにすぎないということが認識できよう。このあたりに，いわゆる数理統計を教えるのと，教育・心理統計を教えるのとの間に大きな違いがある気がする。なお，筆者自身の統計教育に対する考え方については，野口（1983）に簡単に述べたことがあるのでご参照いただければ幸いである。ただし，その後の筆者自身の環境の変化もあり，現在の考え方と必ずしも全てが一致している訳ではない事をご了承いただきたい。

最後に，余談であるが，バッハ家には音楽家，ベルヌイ家には自然科学者という際立った“因子”があるように思うが如何なものであろうか。

4　終わりに

これら3つの論文は色々な意味で大変刺激になった。特に，異分野の研究者が“テスト”に関心を持って研究されている事を知ることができ（知らなかったというのは，筆者が勉強不足なのかもしれないが），幸いであった。今回は十分に討論できなかったが，今後機会があれば，今回の論文の内容に限らずもう少し広く“テスト”研究について話し合ってみたい気がする。

それはともかく，企画者は“実用的な観点等から従来の理論を擁護する議論”を筆者に期待された。各論文毎にそれらしき事は述べたが，筆者は必ずしも従来の理論に固執するつもりはない。というのは，テスト研究に関する理論は現実のテストデータの分析・解釈の役に立たねば話にならない訳で，その為にはひとつの理論で事が足りるとは思われないからである。より精緻な理論・モデルが有効な場合もあるし，逆に，簡便な理論・モデルが有効な場合もあるし，現実的な制約も考慮に入れる必要もある。やはり，その状況に応じて理論の有効性は変わって来る。新しい理論・モデルを研究・開発する事は非常に大切であると同時に，現実の状況を目の前にしてその状況ではどの理論・モデルを適用するのが最適であるかを判断できる力量を持つ事もまた大切な事と考える。筆者自身も“非常識”な誤りを犯さぬ

ように心掛けて，これからも勉強し，研究したいと考えている。

文　　献

Gulliksen, H. 1950 Theory of Mental Tests., Wiley, N. Y.
池田　央　1976　統計的方法 I　新曜社.
野口　裕之　1983　コメント――「児童心理学と数量的方法」について――，児童心理学の進歩　1983年版　金子書房，295-303.
東京大学教養学部統計学教室　編　1991　統計学入門，東京大学出版会.