The Annual Report of Educational Psychology in Japan
2010, Vol. 49, 109-119

測定・評価部門

教育心理学研究の明日のために

孫　　媛
(国立情報学研究所)

はじめに

筆者はこの3年間,『教育心理学研究』の常任編集委員を務め，投稿論文の審査過程に立ち会う機会を得た。その中で，(優れた論文が数々投稿されていたのはもちろんだが)統計学を「手持ちのデータから都合よく尤もらしい結論を導き出す便利なツール」としてしか認識していないように思われる論文に幾度となく遭遇した。例えば，本来は研究者自身が構築した因果モデルを検証するために使われるとされる共分散構造分析の手法を，最初は何のモデルも提示せずに，あれこれとパスを引いて最も高い適合度を得た後で，それをモデルと称して解釈をするといった類の研究が(掲載に至らなかったものも含めて)数多く投稿されていた。

得られたデータを客観的な手続きを通して要約するというのは確かに統計学の大切な役割の一つではあろうが，個々の研究の枠内で閉じるのではなく一般性，普遍性の高い結論を追求するために統計学を活かす，そんな研究がもっとあってもよいのではないかと感じた3年間だった。教育統計あるいは心理測定の領域は，そのような広い見通しに立った方法論を既にいくつも持っている。本稿では，効果量(メタ分析，検定力分析)と項目反応理論に注目して，最近の動向を概観することから始めたい。効果量も項目反応理論も目新しいものではないが，当初は日本だけでなく(この領域の先進国である)米国でも一部の注目にとどまった後，米国で飛躍的に利用が拡大し，日本でも今後の発展が期待されるものである。

Ⅰ．効果量

1．効果量に目を向けることの大切さ

教育心理学研究に限らず心理学研究では，検定を用いることが常識になっている(尾見・川野, 1994)。しかし一方で，検定の偏重・誤用に対する批判も，1940年代頃から現在に至るまで繰り返されている(例えば，Morrison & Henkel, 1970 ; Oakes, 1986)。「帰無仮説が誤りであるとしても，その程度を示せない」というのも，多くの論者が指摘する検定の制約の一つである(Bakan, 1966 ; Cohen, 1994)。Cohen (1988)は，「帰無仮説が誤りである程度(the *degree* to which the null hypothesis is false)」(強調は原文)として，効果量(effect size)を定義している。つまり，効果量とは「相関係数の検定でいえば，$\rho=0$ でないとして，その相関の大きさ」，「平均値差の検定でいえば，$\mu_1=\mu_2$ でないとして，その差の大きさ」に関心を向けるものである。積率相関係数 r は2変数がともに量的変数の場合の標本効果量と考えることができるので，相関係数の場合は効果量推定値を報告することが既に常識となっているといえる。

しかし，平均値差が問題になる場面では，効果量が報告されることは稀である。例えば，対応のない t 検定の場合，最も代表的な効果量指標は標準化平均値差

$$\delta=\frac{\mu_1-\mu_2}{\sigma}$$

であるが，教育統計・心理統計の教育において，この指標に触れることが常識とはいえないだろう(この標準化平均値差を指して効果量と呼ぶ場合もある)。日本の教科書では，1990年に芝・南風原(1990)が「効果量を用いれば，単位の異なる変数を用いた研究の間でも，実験条件の効果の大きさを互いに比較することができる」として紹介している。Cohen の d と呼ばれることもある標準化平均値差の推定値は，

$$d=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{n_1s_1^2+n_2s_2^2}{n_1+n_2-2}}}$$

で求められる(添字付きの $\bar{x}$, s^2, n は，2群それぞれの平均，分散，サンプルサイズ)。

ところで，特に平均値差の検定をする場面において，p 値が0に近いことをもって差の大きさが示されたと解釈する者が後を絶たない。しかし，t 統計量が $t=d\times\sqrt{n_1n_2/(n_1+n_2)}$ と書けるように，検定統計量は一般に，

検定統計量＝効果量×サンプルサイズ

の形で表現できることを知れば，効果量が小さくても(0でない限り)，サンプルサイズが大きければ p 値は0に近づくことを明瞭に理解できるだろう。検定統計量がこの形で表現されることは，米国では遅くとも1990年頃の教科書には，重要な式として記述されていた(Maxwell &

Delaney, 1990 ; Rosenthal & Rosnow, 1991) が，日本では筆者の知る限り，南風原(2002a)が初期のものである。今年度に出た豊田 (2009) では「検定の正体ともいっていい大切な式」(p. 21) として言及されている。積率相関係数 r を論文中に記すのが常識であるのと同じように，平均値差の検定の場面でも効果量推定値を記載するようになれば，p 値に基づいて差の大きさを解釈するという悪弊は減っていくだろう。

検定における「有意」「有意でない」という二者択一の結論はとても便利ではあるが，「有意」という結論だけでは効果の大きさを知ることができないこと，本当は差があるのに「有意でない」という結論が得られる（第2種の誤りの）確率がとても大きなものになっているかもしれない（知らぬが仏）ことにもっと注意が向けられるべきだろう。個々の研究で報告される効果量はあくまでも推定値に過ぎないが，各研究が効果量を（できれば，その信頼区間も）報告することの意味は小さくないと思われる。米国心理学会 (American Psychological Association) は *Publication Manual* (American Psychological Association, 1994, 2001, 2009)で効果量の報告を勧めており，米国心理学会が招集した「統計的推論に関する特別作業班 (Task Force on Statistical Inference : TFSI)」は，「すべての研究が効果量とその信頼区間を報告すべきである(そうすることで，研究を読む者が，多くの標本，計画，分析の安定性を評価できるし，将来の研究に向けて検定力分析，メタ分析への情報が提供される)」と強く勧告している (Wilkinson & Task Force on Statistical Inference, 1999)。試しに *Journal of Educational Psychology* の第101巻1号(2009)を調べたところ，掲載された16論文のうち10論文がANOVAやMANOVAなど群間の平均値差の検定を行っており，それらのすべてが効果量を記載していた。

井上・孫 (2006a, 2006b, 2007) は教育心理学会などで効果量を報告することの意義を訴えてきたが，これまでのところ日本における動きは鈍い。しかし，全く兆しがないわけではなく，今年度に教育心理学会に発表された論文では，及川・及川・青林 (2009) が効果量を記載している。及川らは海外の研究で報告された効果量との比較に言及しており，研究の国際化という観点からも，教育心理学研究において効果量を記載する慣習が広まることを期待したい。上に触れた *Journal of Educational Psychology* の諸論文では，2群の比較ではCohenの d を，3群以上の比較では相関比の2乗 (η^2) あるいは偏相関比の2乗 (η_p^2) を報告するものが大多数であった。論文中における典型的な記載法は，$F(1, 28)=8.64$, $p<.01$, $\eta_p^2=.24$ のようなもので，従来の表記法と大差ない。Cohenの d は t 統計量とサンプルサイズから簡単に計算できるし，偏相関比の2乗はSPSSの一般線型プロシジャで出力することができる。効果量の信頼区間を求めるには特別のプログラムが必要だが，効果量を報告するだけであれば，日本の研究者あるいは学部生，大学院生にとってもそう敷居の高いものではない。今回調べた *Journal of Educational Psychology* の論文でも信頼区間は記されていなかった。

2．メタ分析

検定の限界を補うものとして効果量を報告することを奨励する論者はかなり以前から存在したが，実は米国でも論文中に効果量を記載する慣習はなかなか定着しなかった。米国で効果量が広く注目され，その重要性が認知されるようになった契機はメタ分析の登場だと思われる。1977年のSmithとGlassのメタ分析は，心理療法の効果に関する375件の検定結果を効果量に換算した上で併合し，個々の研究では得られなかった一般性の高い結論を導いた (Smith & Glass, 1977)。1970代後半から80年代にかけて，影響力の大きいメタ分析研究が相次いで発表され，欧米ではメタ分析が急速に広まった。代表的なレビュー誌 *Psychological Bulletin* の掲載論文も近年ではその大半をメタ分析が占めている。これ以降，効果量そのものの重要性も認知（再認識）され，先に記したように，*APA Publication Manual* やTFSIの勧告につながった。一方，日本における心理学分野でのメタ分析の実践例は極めて少ない(孫・井上, 2006)。英語文献のメタ分析がほとんどだという日本の研究者には不利な状況はあるが，メタ分析あるいは効果量の重要性が日本の研究者や学部生，大学院生に浸透しないままなのは惜しい。

3．検定力分析

効果量はそれ自体が重要な意味を持ち，メタ分析のための不可欠のピースであるが，効果量の役割はそれだけにとどまらない。検定は有意水準（第1種の誤りの確率）を定めなければ始めることさえできないが，検定力（あるいは第2種の誤りの確率）は全く考慮しなくても検定ができる。しかしCohen (1988) も説くように，検定力分析の重要性は明らかである。検定力，効果量，サンプルサイズ，有意水準の4項の間には，そのうちの3つを決めれば残り1つが求められるという関係がある。したがって，4種類の検定力分析を考えることができる (Cohen, 1988)。

①有意水準，効果量，サンプルサイズを決めて，検定力を求める
②効果量，有意水準，検定力を決めて，サンプルサイズを求める
③有意水準，サンプルサイズ，検定力を決めて，効果量を求める
④サンプルサイズ，検定力，効果量を決めて，有意水

準を求める

日本では，杉澤 (1999) が1992年から96年の間に発行された『教育心理学研究』に掲載された論文について，対象論文の60%では中程度の効果量を検出できる確率が0.8に満たないという試算を示して検定力分析の重要性を指摘しているほか，村井 (2006) も書籍の中でその重要性を強調している。しかし，これらは例外的であり，重要であることの意識が浸透しているとはいえない。

検定力分析が普及しない理由は数多くあるだろう。概念的な難しさ・わかりにくさ，必要性が認知されていないこと，検定力の計算には特別のソフトウェアが必要であることなど，様々な障害がある。このうち，計算の部分については，南風原 (2002a) がGPOWER (Erdfelder, Faul, & Buchner, 1996) を，村井 (2006) がGPOWERを含む何種類かのソフトウェアを紹介している。GPOWERは比較的使いやすいフリーソフトだが，それでもこれまであまり使われてこなかった。

そんな状況を変えるかもしれないと期待させてくれるのが豊田 (2009) である。豊田は上の4つの検定力分析のうち重要性の高い①②について，その使い方の枠組みを魅力的なネーミングとともに提示している。

事前の分析：上記の②に相当。3通りの効果量（小0.2，中0.5，大0.8）のそれぞれについて，目標とする検定力（例えば，.85や.95など）および有意水準を決めて，研究に必要なサンプルサイズを見積もる分析である。研究を計画する段階で，サンプルサイズの目安を知るために有益である。

事後の分析：上記の①に相当。研究で得られた標本効果量を効果量推定値として使い，自分の行った研究について検定力を知るための分析。期待に反して有意な結果が得られなかった場合などに，研究計画を見直すための分析として有益である。

明日への分析：上記の②に相当。効果量として大・中・小の見込みの値を用いる代わりに研究で得られた効果量推定値を用いる点で，「事前の分析」と異なる。

いろいろな研究領域・テーマごとに，効果量の知見を蓄積していくことが，未来の研究において重要な意味を持つことがわかるだろう。そして，このことは，個々の研究における有意，有意でないという結論を超えて多数の証拠を集めることで，より一般的な結論を導こうというメタ分析の発想ともつながる。過去に行われた同種の研究で報告された効果量やメタ分析で得られた知見を活用することで，将来の研究において必要なサンプルサイズをより正確に自信を持って見積もることができる。

豊田はさらに，『教育心理学研究』などに掲載されたいくつもの論文について，そこで得られた効果量推定値を用いて「事後の分析」「明日への分析」を行った例を多数掲載しているが，これらの結果を眺めるだけでも興味深く，それだけのためにでも，この本を手にとってもらいたい。また，この本では統計解析パッケージRを用いて検定力分析を行う方法も解説されており，Rを使う気さえあれば，自分で検定力分析を行うことも容易である。なお，豊田らは今年度の日本心理学会第73大会において，「やはり，検定力分析はすべきです！」と題するワークショップを企画・主催している（川端, 2009）。

日本教育心理学会第51会総会においても，「教育心理学研究における統計的検定の再考」（森・村井・白川・深谷, 2009）と題する自主シンポジウムが企画され，「統計的検定の問題点と適用上の留意点あれこれ」（吉田寿夫），「非劣性・同等性を積極的に言う方法」（石井秀宗）とともに「検定力検定の実践と検定力」（杉澤武俊）という論題で話題提供が行われている。

II．項目反応理論

1．日本における項目反応理論研究の動向

項目反応理論（Item Response Theory：IRT）については，『教育心理学年報』の測定・評価部門でこれまでに何度も取り上げられている（例えば，速水, 1986；野口, 1989；村上, 1990；中村, 1999；廣瀬, 2004）。項目反応理論に関する研究数は必ずしも多いとはいえないが，この理論の重要性が少なくとも測定・評価部門の執筆者の間では認知されてきたということだろう。そこで，日本における項目反応理論研究の動向を見るために，国立情報学研究所が提供する学術論文データベースCiNiiを用い「項目反応理論」「項目応答理論」「item response theory」をキーワードとする論文をOR検索した結果を，年×掲載場所別に集計した。1983年から2009年までの27年間の累計が10件を超えた掲載元が5誌あり，これらはすべて，日本教育心理学会，日本行動計量学会，日本テスト学会のいずれかに関わるものであった。Table 1に，この3学会に関わる論文件数を年別にまとめた（行動計量学会での発表要旨が『行動計量学』に紹介されたものも検索されたが，これは件数から除いてある）。

『教育心理学研究』に限ってみるならば，絶え間なく論文が発表されていること，1990年頃に1つのピークがあったことがわかる。1989年度の『年報』で村上 (1990) 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」について詳しく紹介している。このシンポジウムに指定討論者の一人として参加した村上が「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

Table 1 日本における項目反応理論研究の動向

年	できごと	『教育心理学研究』	『日本教育心理学会総会発表論文集』	『行動計量学』	*Behavior-metrika*	『日本行動計量学会大会発表論文抄録集』	『日本テスト学会誌』
1983							
1984		1					
1985							
1986							
1987		1					
1988		1					
1989	日本教育心理学会総会自主シンポジウム「項目特性理論の展開」	2					
1990		5					
1991	芝祐順『項目反応理論』刊行	1		1	1		
1992		1				1	
1993		2	2				
1994	池田央『現代テスト理論』刊行	1					
1995		1			1		
1996		1	2		5	2	
1997			3				
1998		1	1				
1999			1				
2000		1	3				
2001	CASEC運用開始	1	3		1	2	
2002	日本留学試験開始,豊田秀樹『項目反応理論』刊行	1			1	1	
2003	日本テスト学会発足	1			1	1	
2004		1		2		5	
2005					2		4
2006	医療系大学間共用試験			1	3	1	2
2007		1		1	1	3	1
2008		1	1			2	3
2009		1		1			3
	総 計	26	16	6	16	18	13

に，少々驚きを禁じえなかった」と述べていることから，この時点で項目反応理論に関心を向ける研究者が多くいたこと，しかしそのことが広くは知られていなかったことが窺える。また，『項目反応理論―基礎と応用―』(芝, 1991)の文献リストを見ると，1978年に東京大学教育学部研究紀要に発表された「語彙理解尺度作成の試み」(芝, 1978)以降1990年頃にかけて，語彙理解力尺度に関連した研究（等化，適応型テストなど）が，『教育心理学研究』および『東京大学教育学部紀要』などに数多く発表されていたことがわかる。

1990年代も『教育心理学研究』が中心であるが，中村は1998年度の『年報』で，「この約10年間の研究の多くは，測定領域を専門とする研究者のもので，方法論的研究に偏っていた」(中村, 1999) とまとめている。教育心理学会総会での発表は，この時期に集中している。

2000年以降になって論文数が増える徴候があるが，この時期は，いろいろな大規模テストへのIRTの導入が進んだ時期でもある。例えば，教育測定研究所のCASECが2001年，日本留学試験が2002年，医療系大学間共用試験が2006年から，項目反応理論に基づいてテストを運用している（日本語能力試験でも2010年から項目反応理論を採用するほか，大学入試センター試験，法科大学院適性試験などが項目反応理論を用いたデータ分析を報告している)。この10年間の論文数の増加は，日本行動計量学会あるいは日本テスト学会の寄与による。特に2003年に設立された日本テスト学会の『日本テスト学会誌』には項目反応理論を用いたテスト研究が数多く掲載されている。

項目反応理論では「特性値 θ を固定したとき，各項目への反応（正答・誤答）は互いに独立である」という局所独立性が仮定されるが，現実のテストではこれが満たされないと思われる場面が多い。佐野(2009)は，SLD(Surface Local Dependence) と名付けられた局所依存構造が存在するとき識別力パラメータが過大推定されることを示し，相互情報量に基づく過大推定検出法の有効性を確認して

いる。

教育課程に基づいた学習の実現状況を調べようとする場合，まず学校（学級）を集団として抽出し，次いで抽出された集団から児童生徒を個人として抽出するという2段抽出法が用いられることが多い。こうして得られたデータは階層性を持ち，また集団ごとに履修状況（履修済みか未履修か）が異なる可能性が大きい。萩原（2009）は，この種のデータに対する2段の項目反応モデルの適用可能性を検討している。

手持ちのデータに項目反応理論を適用するには（プログラムを自作するのでなければ）専用のソフトウェアが必要だが，これまで多くの研究で利用されてきた定評あるソフトウェア（例えばBILOG-MG）は，海外製で入手が面倒だったり，GUIを備えておらず自分でコマンドを記述しなくてはいけないなど使い方が難しかった。熊谷（2009）は，操作しやすいGUIを備えた自作のフリーソフトウェアについて報告している。項目反応理論に興味はあるが敷居が高いと思っていた人には朗報だろう。

Table 1において『日本テスト学会誌』2009年度掲載としてカウントされた論文は以上の3論文だが，以下の2論文も項目反応理論に関連している。

張（2009b）は，問題文が同一で形式が異なる設問（多肢選択形式と穴埋め形式）を同じ受検者集団に解答してもらって得られたデータを，様々な観点から分析することで項目形式の影響を探り，テスト作成に対する貴重な示唆を得ている。張は『行動計量学』でも項目反応理論に基づく研究を発表している（張，2009a）。

野上（2009）は，2001年に運用が開始され多くの人が受検してきたCASECの項目について，項目の利用頻度や繰返し受検が項目の正答しやすさおよび受検者の能力推定に及ぼす影響を，実データとシミュレーションデータの分析を通じて調べている。CASECのように受検者の多いCAT（コンピュータ適応型テスト）におけるメンテナンスの重要性を再認識させてくれる。

『日本テスト学会誌』の各論文はどれも，項目反応理論が日本においても実用化のフェーズに入ったことを感じさせるものであった。

『教育心理学研究』では，高橋・中村（2009）が，学童期の子どもを対象とする語彙・漢字の検査をコンピュータ適応型テストとして開発した過程について報告している。学童の言語能力の尺度化という教育心理学の伝統的なテーマに沿った，『教育心理学研究』にふさわしい研究である。ただ，「語彙」「項目反応理論」といえば直ちに想起される芝の語彙理解尺度と比べて今回の語彙項目プールがどんな特徴を持っているのかに言及してもらいたかった。

国内の項目反応理論研究はこの10年間で活性化しており，若い研究者の活躍も目立つ。それでも諸外国における研究の活況を見るとき（例えば，2009年12月13〜15日に開催された第25回IRTワークショップ〈http://www.utwente.nl/projecten/irtworkshop/〉），研究者の層がまだまだ薄いと感じられる。発表の主要な舞台は『教育心理学研究』から『日本テスト学会誌』に移った感があるが，日本行動計量学会も含めて，関連学会の研究者が相互に刺激し合い，こうした領域に興味を向ける研究者が増えることを期待したい。

本節の最後に，今年度に刊行された図書の中からテストに関するもの2冊を挙げておきたい。『eテスティング』（植野・永岡，2009）には，項目反応理論をはじめとするテスト技術の最新情報や項目反応理論適用の実例などが多角的に紹介されており，研究の現状を最小の努力で正確に知ることができる。一方，労力は覚悟の上で，テストに関わる最前線の動向を包括的に知りたいという人向けには，「発達した最新技術と考え方による公平妥当なテスト作成・実施・利用のすべて」という副題のついた『テスト作成ハンドブック』（Downing & Haladyna, 2006　池田（監訳）　2008）がある。

2．学力調査と項目反応理論

学力調査について2005年度の『年報』で山森（2006）が多くのページを割いているが，その後も，2006年に経済協力開発機構（OECD）による国際学習到達度調査（PISA）が，2007年に文部科学省による全国学力・学習状況調査が行われた。この種の大規模な学力調査といえば想起されるのが米国の全国学力調査（National Assessment of Educational Progress：NAEP）である（荒井・倉元，2008；村木，2006）。米国の児童生徒（第4，第8学年，のちに第12学年も）の全国的標本に対して，いくつかの教科科目の教育達成度を把握するために1969年に始められたNAEPでは，1984年になって項目反応理論を導入している。この種の学力調査に項目反応理論を適用することには，どんな意味があるのだろうか。

(i) 各教科について，児童生徒が何を知り何ができるのかを包括的に把握するためには，教科領域を広くカバーする多数の項目を用いることが望ましい。しかし，一人の児童生徒に過度の負担はかけられない。NAEPでは，マトリックス標本抽出法という被調査者と項目の両方を標本抽出する方法によって，個々の被調査者の負担を小さく抑えつつ広範囲の学力を正確に測ることを可能にしている。得られたデータは全員が同じ項目を受検しているわけではないが，項目反応理論を用いることで全項目の特性が共通能力尺度上で表現され，異なる項目を受検した被験者

同士を同一尺度上で評価することができる。

(ii) NAEPは調査時点における全国的な学力傾向を調べるための主調査の他に，長期的な学力変化を調べるための動向調査を含んでいる。動向調査では複数年にわたり同一の項目を用いており，項目反応理論の適用によって得られた共通尺度上で，学力の経年変化を客観的に追跡することができる。ちなみに，PISAも各領域の尺度化に項目反応理論を駆使しており，2000年，2003年，2006年の年度比較にもその技術を活用している。

2009年末に世間の話題をさらった事業仕分けで，全国学力・学習状況調査も予算削減の対象として俎上に載せられた。だが，全数(悉皆)調査か標本調査かという問題は，予算削減云々とは別に調査の目的という観点から論じられるべきものだろう。文部科学省は，全国学力・学習状況調査の目的として，

(ア) 国が全国的な義務教育の機会均等とその水準の維持向上の観点から各地域における児童生徒の学力や学習状況をきめ細かく把握・分析することにより，教育および教育施策の成果と課題を検証し，その改善を図る，

(イ) 各教育委員会，学校等が全国的な状況との関係において自らの教育および教育施策の成果と課題を把握し，その改善を図るとともに，そのような取組みを通じて教育に関する継続的な検証改善サイクルを確立する，

(ウ) 各学校が各児童生徒の学力や学習状況を把握し，児童生徒への教育指導や学習状況の改善等に役立てる，

を挙げている(文部科学省, 2009)。(ウ)のためには悉皆調査が必要かもしれないが，(ア)(イ)の目的に関していえば，有効な情報を得るための最適な方法はNAEP流のマトリックス標本調査法および項目反応理論の適用に基づくものだと思われる。また，長い間話題になりながらこれまでデータの裏付けがなかった学力低下について生産的な議論を展開するためにも，緻密に計画を練った上でのデータ収集が必要である。今後，全国学力調査の新しい方向性について検討が進められることを期待したい。

III.『教育心理学研究』掲載論文の概観(因子分析を中心に)

過去の『教育心理学年報』の中で，栗田(2007)は1996年度と2006年度，小泉(2009)は2008年度について，『教育心理学研究』掲載論文における統計手法の利用頻度を表にまとめている。これらと比較するために，2008年7月から2009年6月に発行された『教育心理学研究』(第56巻第3号から第57巻第2号まで。以下2009年度)に掲載された38論文(「展望」を除く)で用いられた統計手法のうち，因子分析，共分散構造分析，分散分析の利用頻度を整理した(Table 2)。また，栗田(2007)は，探索的因子分析を用いた研究についてさらに詳しく，それらが採用した初期解と回転法をまとめているので，これに2009年度分を加えてTable 3とした。以下，因子分析について，その使われ方を細かく見ることにする。

Table 2 因子分析・共分散構造分析・分散分析の利用頻度

年度	1996	2006	2008	2009
探索的因子分析	17	20	17	15
確認的因子分析	0	2	6	5
共分散構造分析	1	5	5	5
分散分析	28	25	19	20
論文総数*	46	48	46	38

*2008，2009年度については，論文総数に「展望」を含まない．

Table 3 探索的因子分析の各手法の利用頻度

	年度	1996	2006	2009
初期解	主因子法	14	16	7
	主成分法	3	1	2
	最小二乗法	0	2	2
	最尤法	0	1	3
	ミンレス法	0	0	1
回転法	バリマックス	11	8	1
	プロクラステス	1	1	0
	プロマックス	5	10	13
	オブリミン	0	1	0
	なし(1因子)	0	0	1

1．探索的因子分析

因子パターンの推定法は，かつての主因子法一辺倒から最尤法が増える兆しが見られる。繁桝・柳井・森(2008)は，統計ユーザーの様々な疑問に解答を与えてくれる良書だが，第1版(1999)から「最小2乗法を使えるときは主因子法を使う必要はない」という記述が見える。1996年度の『年報』の中で豊田(1997)も「少なくとも今現在では最小2乗法・一般化最小2乗法・最尤推定法の中から推定法を選ぶのが定石といって良いのではなかろうか」と述べている。10年以上前からいわれているわりには，主因子法からの転換は進んでいないが，主因子法よりも最小2乗法，最尤法が良いという理由が初心者(というより一部の例外を除く大多数の研究者)には伝わらず，慣習の力が働いているということだろう。

回転法に関しては，プロマックス回転の一人勝ち状態に移行している。柳井(2000)は，1999年度の『教育心理学研究』『心理学研究』に掲載された19編の因子分析研究のうち，バリマックス回転13，プロマックス回転4，オ

ブリミン回転1，斜交プロクラステス回転1であったことを報告しており，この時点ではまだバリマックス回転が大半を占めていたが，21世紀に入り，回転法の主役がプロマックス回転に交替したことは明らかである。

因子間の相関に関心を向けるならば斜交解を用いるのが当然だが，因子間の相関に注目せず変数を分類して尺度得点を求めることだけが目的ならば斜交解独自の出力は不要であり，直交解で十分だとも思える。しかし，明快な単純構造が得られないデータでも斜交回転によって単純構造が明確になることがあり，このようなデータに対しては，やはり斜交モデルを選ぶのが自然なのだろう。柳井 (2000) は，最初に斜交回転を行い，因子間相関が低い場合に限って直交回転を選ぶべきだという考えを披露している。

今年度の論文のうちプロマックス回転を選んだ論文すべてで，ある程度の因子間相関が見られ，唯一バリマックス回転を選んだ論文では2つ以上の因子で負荷の高い変数が目立ち，しかも自身の関連する先行研究ではプロマックス回転を用いており，バリマックス回転を選んだ理由が不明である。今年度の研究についてはすべてが斜交回転でもおかしくなかったということであろう。また，今年度は斜交回転としてプロマックス回転だけが選ばれていたが，例えばオブリミン回転の方が合うようなデータはないのだろうか。たいていの統計ソフトでは直交・斜交とも何種類かの回転法を選べるはずで，それぞれの回転法についてどんな特徴があるのか，専門家の教えを請いたいところである。なお，高橋・中村 (2009) における因子分析は，項目反応理論を適用する前提として1次元性を確認するためのもので，当然，回転も行っていない。

因子分析前後の処理の記述に関しては，項目の天井効果や床効果などの点検についてきちんと書かれており，分析後の尺度の信頼性について α 係数だけでなく再検査信頼性も示すなど尺度化の手順が手堅く記述された論文が多かった。また，『教育心理学ハンドブック』中で村上 (2003) が収束的証拠と弁別的証拠を解説した頃は，1つか2つの他尺度との相関係数を示して「妥当性が確認された」と書くような論文が多かったように記憶するが，今年度は，作成した尺度と相関することが予想される既成の尺度，相関しないと予想される既成の尺度それぞれ多数との相関を丁寧に調べる研究も見られた（鈴木・木野, 2008；加藤・谷口, 2009 など）。

他方，変数選択については，相変わらず負荷量の小さい変数，複数の因子に大きな負荷を示す変数を機械的に捨てる（「小さい」「大きい」の数値は恣意的で，研究ごとにばらばら）といった手順が大半だった。変数の増減により因子構造そのものが変化しうることを考えると，長濱・安永・関田・甲原 (2009) のように，ステップワイズ変数選択のような理論的にしっかりとした方法を使う論文がもう少し増えてもよいのではないかと感じた。

2．確認的因子分析

2006年度（栗田, 2007）と比べて，この1，2年で確認的因子分析を用いる論文は明らかに増えている。昨年度の『年報』で小泉 (2009) は，探索的因子分析 (Exploratory Factor Analysis : EFA) を使った17論文のうち6論文が確認的因子分析 (Confirmatory Factor Analysis : CFA) を行っていたことを報告している。今年度はEFAを行ったのちCFAで適合度を確認する論文の他に，2因子を想定して行ったEFAの結果1因子解が適切に思われたのでCFAで適合度を確認したもの（小山, 2009）や，先行研究と同じ因子構造を期待して最初からCFAを行った研究があった（葉山・桜井, 2008；川島・眞榮城・菅原・酒井・伊藤, 2008）。これらのうち葉山・桜井 (2008) のCFAの結果は十分な適合度とはいえない(GFI＝.85, AGFI＝.82, RMSEA＝.06)。因子分析の利用法として，CFAで適合度を検討するというタイプの論文が一層増えそうな気配であるが，南風原(2002b)が指摘するように，適合度の高さは必ずしも尺度の一般化可能性，妥当性の高さを保証しないように思われる。ルーチン的にCFAによる高い適合度を求めるばかりでなく，EFAの積極的な意味を改めて考える時期に来ているのかもしれない。

3．有意傾向とは何か？

ところで，以前から気になっており，今年度の『教育心理学研究』でも多く見られた表現が，p 値が10%未満であるときの「有意傾向が見られる」という言い方である。検定は本質的に白か黒か（有意か有意でないか）の二分法であり，有意傾向というグレーゾーンは存在しないはずである。「傾向」という接尾辞が，第1種の誤りの確率を大きく設定しているという警告ならばよいのだが，差の大きさへの言及（有意だけれど目立たない小さな差）という誤解の温床となる危惧がある。実際，多くの学生が有意傾向は差の小ささを示すと誤解をし，そのようなつもりで論文中に使う傾向があるように見える。

鋤柄 (2002) もいうように，論文における記述は，心理学を学ぶ学生にとってのお手本としての役割を持つ。「有意傾向」という表現は，統計の授業で教えられる検定の説明と矛盾するように思う。p 値を記述するのは常識なのだから，単に「有意 ($p<.10$)」と書けばよいのではないか。第1種の誤りの確率を10%に見込むことを躊躇するならば，結果セクションの最初にでも「この論文では有意水準を10%に設定しない」と宣言した上で，p 値が5%を超える場合は「有意でない」と書くべきではない

だろうか。

IV. 統計学に対する理解促進

研究において統計学が十分に活用されるようになるためには，研究者あるいは研究者の卵たちが統計学へ関心を向け，理解を深めていくことが前提になる。教育心理学の典型的な学習者は数学があまり得意ではなく，方法論にあまり興味を持っていない。彼らが方法論の大切さに気付き，統計学の学習を進めていくためには，それを支援する取組みが必要だろう。

日本教育心理学会第51回総会において，山田・村井・杉澤・寺尾 (2009) は，「文系学生に対する心理統計教育—統計ソフトウェアからみた教育実践—」という自主シンポジウムを企画している。山田らは第49，50回総会でも，文系学生に対する心理統計教育に関する自主シンポジウムを企画しており，継続的な活動に敬意を表したい。今回のシンポジウムでは，3人の指定討論者がそれぞれExcel，R，SPSSを用いた授業について紹介している。これらのソフトウェアはいずれもデータ解析の有力な道具であるが，論文執筆の際だけでなく，それ以前の段階で，実際のデータをいろいろな角度から分析し，得られた結果を注意深く検討する経験を積むことが統計への理解を促進する上でとても役に立つ。特に，Rはデータ解析の学習ツールとしても大きな可能性を持っていると思う。山田らも指摘するようにコマンドラインの操作は障壁にはなるが，それを厭わなければ，意欲的な学生にとって非常に強力な味方になってくれる。

2006年度の『年報』で栗田 (2007) がRへの注目と普及への期待を表明しているが，解説書の類も着実に増えており，本年度も初学者から使えるR本がいくつか刊行されている。その中で特にお薦めしたいのが青木 (2009) である。これまでもWeb上で公開されRを使う人の間で有名だったテキストが書籍になり，参照しやすくなった。Rを使いこなす上で壁になりかねないデータの取扱いが詳しく解説されており，一変量の記述統計から始まり，検定・推定，多様な多変量解析の手法までをカバーする，非常に実用的な良書である。特にグループ別の分析の実例が豊富なのがありがたい。

筆者自身がRでデータ分析をする際によく参照するのが，この本とLigges (2004 石田訳 2006)，間瀬 (2007) である。統計を学びながらRの使い方も覚えようという読者には，山田・杉澤・村井 (2008) が薦められる。Rそのものを解説した書籍以外でも，統計諸手法をRコード付きで解説する書籍も増えている。本年度も，テキストデータ解析 (金, 2009)，データマイニング (豊田, 2008)，(すでに挙げた) 検定力分析 (豊田, 2009) などがRの利用を想定している。Rを習得しておくことから得られる見返りは大きい。

大学生，大学院生に統計を教えていて「何か良い問題集はありませんか？」という質問をされることが多い。需要はかなり見込めるのに，そしてかなり多種類の教科書が出版されているのに，問題集はなぜかあまり見かけない。米国の教科書は分厚く，解説も詳しくて章末に多くの練習問題が付いているのが普通だが，薄い日本の教科書では練習問題が載っていないことが多い。

南風原・平井・杉澤 (2009) は，こうした渇望に応えてくれる。章立ては南風原 (2002a) に準じており，通常の統計の教科書が扱う領域に加えて因子分析と共分散構造分析までをカバーしているが，単に「問題に答えたら終わり」という問題集ではない。各問題に対して解答だけではなく詳しい解説が加えられ，各章末に載せられたトピックは，普通の教科書には触れられていない，だがとても大事な事柄が書かれている (例えば，トピック5-4「検定か区間推定か」を読んでみてほしい)。お手軽ではなく実は本格的な本だということを認識して本気で取り組めば，統計学に対する理解が格段に深まりそうだ (なお，同書の問題やトピックの一部でもRが用いられている)。

おわりに

中国北京師範大学で2009年11月12日から開催された教育測定・評価および統計学学科設置に関する国際シンポジウムに参加した。中国では，近年，政府がテストの品質管理を重要視し，大きな予算を計上して研究センターや組織づくりを全土で推進している。それらの中でも最大のものが，2009年4月に創立された北京師範大学の教育統計・測定研究所で，全国学力テスト調査が実施される場合には，項目作成から，実施，分析までを行っている。実践的な課題，予算面のサポートを背景に，海外の専門家の協力を受けながら，テスト理論・測定評価・統計学分野を融合する教育研究の体制が着々とかつ急速に整いつつあり，勢いを強く感じた。米国の研究もバイアス探索の必要性からDIFの研究が隆盛になり，NAEPにおける項目反応理論の採用以降に研究が活性化するなど，課題解決型の研究が多いといわれるが，中国も似た道を進みながら研究が発展しているように見える。項目反応理論の項で見たように，日本でも動きはあるものの，まだその動きは緩やかで小さい。測定・評価に関わる研究者・教育者が協力・連携して，実践・理論の両面から研究を発展させていかねばならないと感じた次第である。

そのような意味で，他学会ではあるが日本テスト学会の充実ぶりを心強く感じた。日本テスト学会は，2003年の発足以降，年次大会や年刊の学会誌，各種の研究会・

ワークショップの開催を通して，テストに関する知見の普及・促進に努めている。2009年度の『日本テスト学会誌』には先に概観した項目反応理論に関わる研究以外にも，テスト・評価に関わる興味深い研究が掲載されている。数理的な研究だけではなく，応用的な研究，教育社会学的な研究も多い。

2009年の日本テスト学会第7回大会（名古屋）では，池田央・柴山直両氏の企画による「テスト研究者がなすべき社会的役割」（池田・柴山, 2009）という公開シンポジウムが開かれ，「テストの専門家の実態とその不在」（木村拓也），「大規模テストにおけるテスト研究者の役割」（柴山直），「新しいテスト技術による取り組み方―医療系大学間共用試験を例に―」（前川眞一），「日本のテストの将来に向けて」（池田央）という話題が提供された。テスト学会はチュートリアルも頻繁に開催している。教育心理学会に所属される方でも，測定・評価に関心を持たれたら，ぜひ参加されることをお勧めしたい。

本稿では日本テスト学会に多くの字数を費やし，測定・評価に比重を置く研究の発表の場が，日本テスト学会に移っていくような印象を与えたかもしれない。しかし，今後の方向性として筆者が期待するのは，小さなパイの奪い合いではなく，パイそのものが大きく充実していくことである。測定・評価に関心を向ける研究者・学生・大学院生が，発表の選択肢が広がったと認識をして意欲的に発表をすることで，この領域の研究が活性化することを願っている。そのことは必ず教育心理学研究一般の質の向上にも資することになるだろう。

長年にわたり日本の教育測定・心理統計を牽引してこられた芝祐順先生が2009年11月15日にご逝去されました。私は先生の最後の指導学生として公私ともに大変お世話になりました。ここに謹んでご冥福をお祈り申し上げます。

引 用 文 献

American Psychological Association (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, DC : Author.

American Psychological Association (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC : Author.

American Psychological Association (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC : Author.

青木繁伸 (2009). Rによる統計解析 オーム社

荒井克弘・倉元直樹 (2008). 全国学力調査日米比較研究 金子書房

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, **66**, 1-29.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ : Lawrence Erlbaum Associates.

Cohen, J. (1994). The earth is round ($p<.05$). *American Psychologist*, **49**, 997-1003.

Downing, S. M., & Haladyna, T. M. (Eds.) (2006). *Handbook of test development*. Mahwah, NJ : Lawrence Erlbaum Associates. (ダウニング, S. M.・ハラダイナ, T. M. 池田 央（監訳） (2008). テスト作成ハンドブック 教育測定研究所)

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER : A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, **28**, 1-11.

南風原朝和 (2002a). 心理統計学の基礎―総合的理解のために― 有斐閣

南風原朝和 (2002b). モデル適合度の目標適合度―観測変数の個数を減らすことの是非を中心に― 行動計量学, **29**, 160-166.

南風原朝和・平井洋子・杉澤武俊 (2009). 心理統計学ワークブック―理解の確認と深化のために― 有斐閣

萩原康仁 (2009). 履修状況を考慮した2段の項目反応モデルとその適用 日本テスト学会誌, **5**, 23-39.

葉山大地・桜井茂男 (2008). 過激な冗談の親和的意図が伝わるという期待の形成プロセスの検討 教育心理学研究, **56**, 523-533.

速水敏彦 (1986). 測定・評価に関する研究の動向 教育心理学年報, **25**, 107-116.

廣瀬英子 (2004). 測定・評価に関する研究動向と展望―テスト研究と評価研究― 教育心理学年報, **43**, 99-106.

池田 央 (1994). 現代テスト理論 朝倉書店

池田 央・柴山 直 (2009). テスト研究者がなすべき社会的役割(公開シンポジウム) 日本テスト学会第7回大会発表論文抄録集, 30-39.

井上俊哉・孫 媛 (2006a). 心理学研究における効果量とその信頼区間 日本行動計量学会第34回大会論文抄録集, 50-51.

井上俊哉・孫 媛 (2006b). 教育心理学研究における効果量について 教心 **48**, 94.

井上俊哉・孫 媛 (2007). 教育心理学研究における効果量報告の意義 教心 **49**, 106.

加藤 司・谷口弘一 (2009). 許し尺度の作成の試み

教育心理学研究, **57**, 158-167.
川端一光（企画）（2009）. やはり検定力分析はすべきです！—power to the people— 日本心理学会第73大会, WS001.
川島亜紀子・眞榮城和美・菅原ますみ・酒井 厚・伊藤教子（2008）. 両親の夫婦間葛藤に対する青年期の子どもの認知と抑うつとの関連 教育心理学研究, **56**, 353-363.
金 明哲（2009）. テキストデータの統計科学入門 岩波書店
小泉令三（2009）. ユーザーから見た測定と評価そして研究法の動向 教育心理学年報, **48**, 123-129.
小山義徳（2009）. 英単語学習方略が英語の文法・語法上のエラー生起に与える影響の検討 教育心理学研究, **57**, 73-85.
熊谷龍一（2009）. 初学者向けの項目反応理論分析プログラムEasyEstimationシリーズの開発 日本テスト学会誌, **5**, 107-118.
栗田佳代子（2007）. 測定・評価に関する研究動向と展望—統計的データ解析法の利用の現状とこれから— 教育心理学年報, **46**, 102-110.
Ligges, U. (2004). *Programmieren mit R*. Heidelberg, Germany : Springer-Verlag. （リゲス, U. 石田基広（訳）（2006）. Rの基礎とプログラミング技法 シュプリンガージャパン）
間瀬 茂（2007）. Rプログラミングマニュアル 数理工学社
Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data : A model comparison perspective*. Pacific Grove, CA : Brooks/Cole.
文部科学省（2009）. 全国学力・学習状況調査の概要〈http://www.mext.go.jp/a_menu/shotou/gakuryoku-chousa/zenkoku/07032809.htm〉(2009年11月30日)
森 敏昭・村井潤一郎・白川佳子・深谷優子（企画）（2009）. 教育心理学研究における統計的検定の再考 教心**51**, 自主シンポジウムE4.
Morrison, D. E., & Henkel, R. E. (Eds.) (1970). *The significance test controversy*. Chicago, IL : Aldine.
村井潤一郎（2006）. サンプルサイズに関する一考察 吉田寿夫（編著） 心理学研究法の新しいかたち（pp. 114-141） 誠信書房
村上 隆（1990）. テスト理論と現実の「はざま」で 教育心理学年報, **29**, 92-100.
村上 隆（2003）. 測定の妥当性 日本教育心理学会（編） 教育心理学ハンドブック（pp. 159-169） 有斐閣
村木英治（2006）. 全米学力（NAEP）概説—テストデザインと統計手法について— 東京大学大学院教育学研究科教育研究創発機構 教育測定・カリキュラム開発（ベネッセコーポレーション）講座2005年度研究活動報告書(2), 51-66.
長濱文与・安永 悟・関田一彦・甲原定房（2009）. 協同作業認識尺度の開発 教育心理学研究, **57**, 24-37.
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.
野上康子（2009）. コンピュータ適応型テストCASECにおける項目の長期使用の影響について 日本テスト学会誌, **5**, 145-164.
野口裕之（1989）. 教育心理学研究に於ける測定・評価に関する研究の動向 教育心理学年報, **28**, 115-124.
Oakes, M. (1986). *Statistical inference : A commentary for the social and behavioural sciences*. Chichester, UK : Wiley.
及川 晴・及川昌典・青林 唯（2009）. 感情誤帰属手続きによる潜在目標の測定—潜在および顕在目標による日常行動の予測— 教育心理学研究, **57**, 192-200.
尾見康博・川野健治（1994）. 心理学における統計的手法再考—数字に対する“期待”と“不安”— 性格心理学研究, **2**, 56-67.
Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research : Methods and data analysis* (2nd ed.). New York : McGraw-Hill.
佐野 真（2009）. 相互情報量を用いた項目識別力の課題推定の検出 日本テスト学会誌, **5**, 3-21.
芝 祐順（1978）. 語彙理解尺度作成の試み 東京大学教育学部紀要, **17**, 47-58.
芝 祐順（1991）. 項目反応理論—基礎と応用— 東京大学出版会
芝 祐順・南風原朝和（1990）. 行動科学における統計解析法 東京大学出版会
繁桝算男・柳井晴夫・森 敏昭（2008）. Q&Aで知る統計データ解析［第2版］ サイエンス社
Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, **32**, 752-760.
杉澤武俊（1999）. 教育心理学研究における統計的検定の検定力 教育心理学研究, **47**, 150-159.
鋤柄増根（2002）. 研究法の理解とデータ分析における学生の誤解 教育心理学年報, **41**, 104-113.
孫 媛・井上俊哉（2006）. 日本におけるメタ分析研究の現状 教心**48**, 95.

鈴木有美・木野和代 (2008). 多次元共感性尺度(MES)の作成—自己指向・他者指向の弁別に焦点を当てて— 教育心理学研究, **56**, 487-497.
高橋 登・中村知靖 (2009). 適応型言語能力検査(ATLAN)の作成とその評価 教育心理学研究, **57**, 201-211.
豊田秀樹 (1997). 測定・評価と共分散構造モデル 教育心理学年報, **36**, 119-127.
豊田秀樹 (2002). 項目反応理論 朝倉書店
豊田秀樹 (2008). データマイニング入門—Rで学ぶ最新データ解析— 東京図書
豊田秀樹 (2009). 検定力分析—Rで学ぶ最新データ解析— 東京図書
植野真臣・永岡慶三 (2009). eテスティング 培風館
Wilkinson, L., & Task Force on Statistical Inference (1999). Statistical methods in psychology journals : Guidelines and explanations. *American Psychologist*, **54**, 594-604.
山田剛史・村井潤一郎・杉澤武俊・寺尾 敦 (企画) (2009). 文系学生に対する心理統計教育—統計ソフトウェアからみた教育実践— 日本教育心理学会第51回総会, 自主シンポジウム F6.
山田剛史・杉澤武俊・村井潤一郎 (2008). Rによるやさしい統計学 オーム社
山森光陽 (2006). 学力低下論争, 目標準拠評価の定着, 学力テストブームの狭間で 教育心理学年報, **45**, 92-103.
柳井晴夫 (2000). 因子分析法の利用をめぐる問題点を中心にして 教育心理学年報, **39**, 96-108.
張 一平 (2009a). 項目形式とその応答法が項目値に及ぼす影響について 日本テスト学会誌, **5**, 53-64.
張 一平 (2009b). 2パラメータと3パラメータ項目反応モデルにおける比較 行動計量学, **36**, 15-24.