

測定・評価部門

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—

廣 瀬 英 子

(東京女子大学現代文化学部)

1. はじめに

本稿は、2002年7月から2003年6月末までに発行された「教育心理学研究」(第50巻第3号から第51巻第2号まで)、「心理学研究」(第73巻第3号から第74巻第2号まで)に掲載された論文、および、同期間に刊行された書籍、そして、2003年度の「日本教育心理学会第45回総会」,「日本心理学会第67回大会」,「日本行動計量学会第31回大会」,「日本テスト学会第1回大会」の各発表論文集に掲載された論文を中心にレビューを行った。2003年3月には日本教育心理学会から待望の「教育心理学ハンドブック」(日本教育心理学会,2003)が刊行されている。そのハンドブックの中にも測定・評価の分野の最近の研究動向についてのレビューの部分があり、測定・評価研究の中心的テーマはテスト理論、心理統計学、教育評価である、とまとめられている。本稿はできればその3点を包括して議論を行いたかったが、筆者の関心の高いテスト研究と評価研究の領域に関連した論文が主になった。本稿であまり触れることのできなかった心理統計学(多変量データ解析法や実験計画法など)のレビューは来年度以降の著者に託したい。なお、上記の範囲で論文を紹介するなかで関連してくる他誌に掲載された論文も、一部ここに含めることにする。

2. 評価に関する研究

2.1 教育と評価

2003年度の日本教育心理学会第45回総会では、森敏昭・深谷優子両氏の企画による「あらためて教育評価を問う」という自主シンポジウム(森・深谷・梶井・村山・鹿毛・大塚・村井, 2003)が持たれたほか、学習評価に関する発表が複数見られた。梶井(2003)は、小学生児童の作文評価のために学習指導要領を反映させて作られた18の評価項目の妥当性を、学年差を識別できる項目であるかどうかという点から検討した。その結果、教員にとって学年差を識別することの難しい評価項目が多いことが示された。夏堀(2003)は、児童の創作した物語を教師が評価する場合の評価基準が、夏堀(2002)の調査で明らかになった一般の社会的な信念(「教師というものは道徳的な物語を高く評価

し、破壊的な物語を低く評価する」など)と合致しているかどうかを検討した。その結果、実際に教師が評価する場合に必ずしも社会的信念通りではないことが確認された。評価基準がどうあるべきかという問題は、教職にたずさわる人にもそうでない人にも経験に基づいた思い入れがあり、まとまりにくいであろうが、このような具体的な場面での評価基準に関する実証研究が着実に蓄積・公開されていくことによって、評価基準に対する見解が統合されていくのではないかと考えられる。

一方、犬塚(2003)は中学2年生を対象とした研究を行い、生徒自身が他者の作文を読むという文章評価活動の効果を検討した。これは相互評価を学習の一環として捉えた研究ということができる。同様に廣瀬(2003)は廣瀬(2002)に続く一連の研究のなかで、大学生に研究発表用のPower Pointのスライドを作成させ、互いに相互評価させる試みを行っている。今後、学習活動に評価プロセスを積極的に組み込んだ形での効果的な学習に関する研究が増えていくことが期待される。

山森(2002, 2003)は中学校英語科の観点別評価(関心・意欲・態度)において、一般化可能性理論(GT: Generalizability Theory)を用いて評定項目の検討を行った。これは実際場面におけるGTの適用として、おそらく日本で最初の研究発表と思われる。なお一般化可能性理論に関しては、体系的なテキスト(Brennan, 2001)が出版されている。

2.2 学生による授業評価

学生による授業評価に関しては大塚(1994)でもすでに言及されているように、決して新しい話題ではない。自主的に独自の授業評価アンケートを自らの担当授業で実施して授業改善に役立ててきた教員は、これまでも少なくなかったであろう。しかし、例えば筆者の勤務先において全学的にすべての科目で学生による授業評価を実施するようになったのが2003年度後期からであることにも伺えるように、学生による授業評価が授業カリキュラムの中に組み込まれるようになるには時間がかかっている。FD(Faculty Development)を受け入れようという考え

方は、ようやく多くの教員に浸透してきたのではないか。

大学による全学的な取り組みの一例として、服部(2003)が報告された。この授業評価尺度には学習意欲、学習の方向性、体系的性、知的刺激、授業参加、評価・提示技術、教材の7因子があげられている。学生による授業評価を授業カリキュラムの一環として実施するとなると、どうすれば学生に対する質問の内容や分量が適切なものになるかを検討せねばならず、どの大学でも似たような作業を必要としているはずである。それぞれの大学が手探りで作成し実施しているならば、教育測定・評価研究の中で検討し、積極的に情報交換を行い、一般化を図っていくことも意義深いことなのではないだろうか。

3. テストに関する研究

3.1 テストに関するシンポジウム

この1年は、テストに関するシンポジウムがさかんに行われた年でもあった。2003年度の日本行動計量学会第31回大会では、野口裕之氏の企画のもとで「日本語教育とテスト」と題したシンポジウムが行われた。日本語能力を測定するテストとしては、「日本語能力試験」が有名であるが(日本国際教育協会と国際交流基金によって世界約40ヶ国で実施され、年に20万人以上が受験している)、Can-do-statements(三枝, 2003)や日本語能力簡易試験 SPOT(小林, 2003)などの、比較的構造が単純ながら実用的なテストについての話題が提供された。また、諸外国で独自に行われている日本語テストもある。和田(2003)によりオーストラリアにおける大学入学資格試験としての日本語能力試験が紹介され、その国の日本語教育事情に合わせたテストの必要性が示された。

2002年11月16-17日には大学入試センターにより、入試問題作成関係者による大学入試に関する国際シンポジウムも開かれた(大学入試センター, 2002)。初日は英語、2日目が数学について、中国、イギリス、アメリカ、韓国、日本の代表者が、それぞれの国における問題作成に関する困難な取り組みについて論じた。聴講した筆者の印象では、入試問題の作成に対する考え方が国によってかなり違うこと、日本はどちらかというとイギリスに近く、他の国はアメリカに近いように思われた。しかし、いずれの国でも問題作成にあたってシラバスを明確化し、作問ミスを起こさないために明確な手続きとチェック体制を確保することに力をいれていることが明らかとなった。なかでもアメリカのETSの問題作成過程は、長年の経験の上にたって詳細なチェックシステムが確立されており、学ぶべきことが多くあったように思われる。

3.2 日本テスト学会の発足

2003年5月には日本テスト学会が発足し、10月4日には第1回大会が繁峙算男実行委員長のもとで東京大学駒場キャンパスで開催された。これまでのテストに関連した研究の蓄積があつてのことである。

学会の設立趣意書(<http://www.jartest.jp>)によると、近年のテストを取り巻く社会的国際的環境の変化、また技術的進歩は著しく、テストとそれがもたらす評価の説明責任は一層大きくなろうとしている。こうした環境の変化に応えるには、テスト研究は一専門分野だけにとどまることなく、教育界はもとより、産業界も含めた広い分野の人たちとの協力によって初めて可能であるとされる。そうした実践家と研究者を含む情報交換と研鑽の場としての新しい学会が設立されたとある。このことから、今までテスト開発と実践に力を尽くしてきた教育心理学がここに大きな役割を果たすと言えるであろう。

柳井晴夫氏による企画セッション「心理テストの過去・現在・未来」(柳井, 2003)では、5人の話題提供者による発表が行われた。知能テストのこの10年間の変化については、前川久男(2003)により、Stanford-Binet-4における表示が、比率IQからWechsler式と同様の偏差IQへの移行がみられたこと、また知能検査の問題内容も大きく変わろうとしていることが報告された。それは認知心理学や認知神経心理学の知見も取り入れたものであり、それらの成果に基づくK-ABCやDN-CASなどの新しいタイプのテストも示された。

また、柏木(2003)により、すでに発表されている日本のいくつかの性格テストをBig Five的観点から分析した結果の整合性が示された。

室山(2003)では、コンピュータを利用した職業相談システムCACGs(Computer Assisted Careers Guidance System)のひとつとして日本で開発された職業適性診断システムIn★Sites2000についての紹介がなされた。CACGsは単に相談者の適性をみるためだけではなく、自己理解、職業理解、意思決定という職業選択のステップを相談者自身の判断で統合的に進められるよう支援するシステムとして、欧米諸国で実用化が進められている。日本でも今後広く利用されるようになることが望ましい。CACGsの利用と評価については、「教育心理学研究」の室山(2002)の原著論文に詳しい。

佐伯・西山・高木・柳井・道場・日野原(2003)によるLPC式生活習慣検査の開発報告は、保健医療領域でのテスト開発の例である。毎日の生活習慣は、その人の健康的体質を維持する上で基本的なことである。その意味でこうした尺度の作成と標準化は日々の生活習慣を管理するのに役立つであろう。LPC式生活習慣検査は健康に関

しての尺度であるが、同じことは児童の学習習慣についても考えられる。毎日の生活態度や勉学態度はその後の児童の発達に影響を及ぼすに違いないが、なかなか眼に見える形で示されるケースは少ないように思う。そうした日々の学習習慣と関係づけられた指標が存在すれば、教育現場での具体的な指導に役立つのではないだろうか。

3.3 テストの等化・標準化の問題

テストに関するトピックスの中で注目されたのが、異なる得点の比較を可能にする等化、そして意味ある得点にする共通尺度化、あわせて標準化と呼ばれる問題である。日本行動計量学会第31回大会ではテスト得点の標準化に関して4件の発表が行われた(石塚, 2003; 前川眞一, 2003a; 村上, 2003b; 椎名, 2003)。また、日本テスト学会大会の公開シンポジウムでも、前川眞一(2003b)がこの問題を取り上げた。前川眞一(2003b)は日本で行われる大規模試験(センター試験など)の特徴として、年1回の同一問題一斉実施、プリテストなしの新作問題(再利用しない)、試験問題の公開、素点の利用などをあげている。これらの特徴は、前年の試験とは関係なく年1回独立に施行される試験を念頭においている限りは大きな問題とはならないかもしれないが、成績が複数年にわたって有効であったり、同一年に複数回実施されるなかでいずれかの成績を利用できたりするとなると事情が変わってくる。素点のままでは、どの受験時のテスト得点を利用するかで受験者にとって不公平が生じるからである。これは、たとえ素点を偏差値に変換しても解決されない。試験問題が異なる上に、受験者集団も異なるからである。

アメリカのTOEFLが大学入学のためには550以上、一流大学では600以上必要などと言われても不自然でないのは、それらが毎年同一基準で比較可能な共通尺度に標準化されているからである。このことにより、年を追って受験生の学力が上昇傾向にあるか下降傾向にあるかについても議論可能となる。

しかし、そのためには新作問題の予備実験や問題の非公開制、問題項目のプール制、問題作成者と編集者の分離独立性といったテストシステムそのものの検討が議論されなければならない。こうした問題は、大学入試センター研究開発部より、「大学入学者選抜における評価の標準化に関する研究」(代表者石塚智一, 2002)として報告書が出ている。そこでは外国の状況、アメリカのACTのACT Assessment Programの学力テスト、ETSのSAT (Scholastic Assessment Test)とTOEFL (Test of English as a Foreign Language)、イスラエルのNITE (National Institute for Testing and Evaluation)による大学入学者選抜のための共通試験PET (Psychometric Entrance Test)の例が報

告されている。得点等化の方法としては、ACTでは無作為等化集団法(random equivalent group design)のもとに集められたテストデータに等百分位法(equi-percentile method)が適用され、ETSでは項目反応理論(IRT)を用いたTrue Score Equatingが用いられている。一方NITEでは古典的テスト理論にもとづくTucker法が用いられている。得点等化法はテスト研究者にとって主要な関心テーマのひとつであり、Angoff(1971), Holland & Rubin(1982), Petersen, Kolen, & Hoover(1989), Kolen & Brennan(1995), von Davier, Thayer, & Holland(2003)など多くのテストが出版されている。提案される方法は多岐にわたるものの、試験に欠くことのできない必須検討事項として外国では共通認識されているようだ。

日本では村上(2003a)によって等化の問題についての研究結果がまとめられた。この報告書はこれからのテスト実践に向けて不可欠な問題についての充実した研究である。また、野口・安藤(2003)と安藤・野口(2003)は項目公開制を採る大規模能力試験における等化法を提案している。今のところ一度実施したテスト項目を一般に公開する文化がある日本では、そのことを含めて等化を行うという現実的な提案も必要とされるためである。日本心理学会第67回大会での荘島(2003a)による等化についての小講演では、聞き手の規模からみて関心の広さが伺え、わかりやすい説明によってこの問題の理解も広がったものと推察される。

しかし、世間一般の関心はまだ高いとはいえない。内田(2002)による調査でも、問題の重要性がそれほど大きく意識されているようには思われない。それはまだ問題が顕在化されていないことや情報不足が原因であるように思われる。

それでも、国内でこうした等化技術が実際に全く使われていないというわけではない。斉田(2003)は県単位で実施される英語能力テストの高校1年生のテスト成績をもとに年次別学力変化の推移の検討を行った。1995年度から2002年度までの8年間にわたる延べ12万人のデータがあるが、問題構成やテスト形式はほぼ共通しており、因子分析の結果、少数の不良項目を除けば各年の因子構造はほぼ同一であることが確かめられている。しかし、実施される試験問題は年ごとに違い、受験者も年ごとに異なっているので、これらを相互に比較する手掛かりがない。そこで、8年間のほぼ中間にあたり、また信頼性係数の値も高かった1999年度の試験問題から28題を選んでアンカーとし、また各年度から18題の問題を選んで7組のテストセットを用意した。これを県下の協力校18校3215名に実施し、共通項目の1999年度問題を手掛かりに、他の年度の問題を項目反応理論を用いてそれに合わせる

形で等化し、その共通尺度値の年次変化を分析した。

その結果、共通尺度値による英語能力分布は年を追うごとに山が左に移動し、平均値が低下していることが示された（標準偏差については目立って大きな変化は見られない）。特に、改定指導要領が実施され、旧指導要領による生徒が全部卒業してなくなった年以降の生徒の成績が目立って低くなり、年ごとにそれは低下しつづけている。しかも英語の授業時間が短縮されたのちの2002年度のデータは前年をさらに大きく下回るものであったが、それが本当に授業時間の短縮によるものかどうかを確かめるには、2003年度以後のデータを待つ必要がある。

こうした項目データをもとに、斉田・服部（2003a）はさらに項目反応理論を応用して、項目プールを利用した難易度別高校英語テストを作成した。それを利用して今度は高校1, 2, 3年生の間の垂直等化を試み、英語の学力変化を分析している（斉田・服部, 2003b）。

このような研究は、アメリカのSAT, TOEFLや大規模な学力調査NAEP (National Assessment of Educational Progress)のように、最初の実施計画を立てる段階から周到に組み込んでおけば、改めて特別な研究を別途実施しなくても済むものであり、学力テストはそうあるべきであろう。学力テストのシステム化が必要である。

大規模テストにおける標準化の重要性は医学教育の分野でも認識され始めている。医学教育における共用試験の試みである（佐藤, 2003; 仁田・奈良・吉田・寺尾・石田・福島・齋藤・福田・高久・麻生, 2003）。医師の養成にあたり、全国の大学で学ぶ医学生者の知識や技量に差があっては具合が悪い。そこで、医師に必須の共通ガイドラインに基づいた全国共通のモデル・コア・カリキュラムを策定し、卒業前臨床実習に入る前の医学生に与えるべき学習内容を標準化し精選化することが求められている。それにとまって、学習結果を評価するための試験に対しても、個別の大学の枠を越えた共通尺度による評価の要請が生まれてきたという。共用試験は平成17年度からのコンピュータによるテストの実施を目指して、すでに2回の試行テストが試みられている。

3.4 テストのCBT化

コンピュータによるテスト (CBT: Computer-Based Testing) の実用化研究は、これからますます加速していくことと思われる。以前、廣瀬（2000）は、パーソナリティ、興味、社会的態度等を測定するための多段階評定法による尺度を中心に、テストのコンピュータ化への動きを展望した。テストのコンピュータ化は当時国内ではまだそれほど一般化されていなかったが、ここ数年のあいだに関心は高まってきているようである。

前出の室山（2002, 2003）があげた職業適性診断システム（In ★ Sites2000）は国内でCBTが実用化された例である。前出の医学共用試験もCBTを取り入れ、実用化に向けての試行テストが続けられている（佐藤, 2003; 仁田ほか, 2003）。ほかにも、適応型テスト (CAT: Computerized Adaptive Testing) の原理を取り入れてインターネット受検を可能にした英語能力テストCASEC (Computer Assessment System for English Communication) の開発過程が林則生（2003）により報告されている。

語学教育では以前からコンピュータを利用した取り組みが盛んである。中でもCALL (Computer Assisted Language Learning) システムを取り入れている学校は多い。ここでその全貌をレビューすることは無理であるが、それとは別に、高度情報通信技術と音声認識を取り入れた電話による英語の聞き取りと発音テストとして原田（2003）によって紹介されたPhonePassは、興味を引くものである。日本語教育の分野では、日本語教育学会の西原鈴子氏らが筆者（廣瀬）とともに新しいCAT形式の日本語テストを開発中である。それは、能力認定機能を重視した現在の日本語能力試験よりも学習者の学習指針となるように、学習奨励としての機能を重視したものとなるであろう。

個人研究では、菊地（2003a, 2003b）はWordやExcelを含むMicrosoft Officeを利用したCATの開発を、また小方（2003）はハイパーグラフによる数式答案の評価法を報告している。これからはCBTによるテスト法の試みは増えてくるであろう。

ただ、こうした技術的発展は常に光と影をともなうものであり、その長短についてはいずれ評価されていくであろう。そのための研究活動が今後期待される。アメリカでは大規模なCBTが実用化されてすでに約10年が経過している。その間、研究と実用化の経験の蓄積も増えた。CBTに関する研究書は次々と出版されている。たとえば、van der Linden & Glas（2000）、Thissen, & Wainer（2001）、Parshall, Spray, Kalohn, & Davey（2002）、Mills, Potenza, Fremmer, & Ward（2002）などがあげられよう。これらは今後さまざまなコンピュータテストを開発する上で参考になる。

さらに最近のCBT研究の動きとして、コンピュータによる問題の自動作成（Irvine & Kyllonen, 2002）や論文の自動採点（Shermis & Burstein, 2003）などの研究も進められ、その一部はすでに実用化されている。日本でも石岡・亀田（2002）のjeraterなどによる作文評価の試みが報告されている。池田（2003）はこれまでのコンピュータを利用した心理・教育テストの動向を概括している。

3.5 項目反応理論とテストへの応用

項目反応理論の理論的研究としては、荘島・豊田(2002)の項目反応理論における Cronbach の α 係数の推定問題、また、実験計画場面への応用として荘島(2003b)がある。日本では多くの試験が初出項目を使わなければならない状況を考慮して、尾崎・豊田(2003)は一対比較 IRT モデルによる未使用項目のパラメタ推定を提案している。また大澤(2003)は Performance Assessment における評定者特性の分析に、多相項目反応モデルに代わる階層評定者モデルの可能性を指摘している。

特異項目機能 DIF (Differential Item Functioning) に関する研究としては、熊谷(2003)と井上・孫(2003)がある。熊谷(2003)は BILOG-MG と SIBTEST を比較している。井上・孫(2003)は、高校15校の英語・国語・数学の学力検査データに対して Mantel-Haenszel (MH) 法による分析を行い、いくつかの項目で学校の違いによる DIF が見られることを報告している。

いくつかの大規模テストに関して、項目反応理論を用いた分析研究も報告されている。大学入試センター試験の「英語」については荘島・吉村・中畝(2003)が経年的変化を考察している。日本語能力試験については、毎年発行される「日本語能力試験分析評価に関する報告書」(日本語教育学会, 2003)に詳しい分析結果がまとめられている。日本語能力試験の問題は1級から4級に分かれており、1級と2級について項目反応理論による分析が示されている。そのほか、項目反応理論によるパーソナリティ・態度テストの開発については、日本心理学会第67回大会のなかで野口裕之・渡辺直登両氏の企画(野口・渡辺, 2003)によるワークショップが開催され、16PF 日本語版や職業性ストレス尺度などへの適用が紹介された。なお、項目反応理論と直接関係はないが、吉村・中畝・荘島・石岡(2003)によると、大学入試センターでは試験統計情報のデータベース化が進められている。このことは今後の入試問題作成環境を整備する上で特記すべきことと思われる。

3.6 テストと学習

最後に、テストは測定、評価、そして学習の全般を総合的に含む存在であると言える。テストと学習の両方を視野にいた研究がいくつか見られた。林篤裕(2003)は、単にテスト解答の正誤情報だけでなく、その問題を解答するために必要な思考プロセスの属性分析を通じて学習達成度を測る Rule Space 法の適用例を報告した。

こうしてテストから得られる情報は、その後の学習指導に効果的に生かされていくことが望まれる。村山(2003)はテスト形式が学習方略に与える影響を検討した。その

結果、授業後に行われる学習内容の確認テストが空所補充型で出題されることが事前にわかっている場合と、記述式テストで出題されることが事前にわかっている場合では学習方略の深さが異なり、また、学習中のノートの書き込み量も異なっていた。学習の内容に合った適切な学習方略を学習者に求めるならば、総まとめとしてのテストをどのような形にするか、その形式もはじめてから熟慮する必要がある。その意味では、学習のプロセスがテストの一部となるようなテストが増えていっても良いのではないかと考えられる。

4. 学部の統計教育とテキスト

テスト関係の研究は、もはや一部の研究者の関心事ではなくなりつつある。全く別のフィールドで活躍している人でも、仕事の必要に迫られてテストの作成や運営・分析に関わることがある。また、将来学校教育や教育ビジネスなどの教育界に進む大学生は、これからも毎年必ずいるであろう。テストに対する基本的な知識や統計的手法を伝える場は今後も広がっていかねばならない。現在筆者は文科系の学部生を対象とした統計学の授業をいくつか担当しているが、私立文系用の受験勉強をして入学し、できれば数字は見たくないという学生に理解してもらうために、毎年、試行錯誤を繰り返している。対象となる学生にちょうど合うようなテキストはないかと、日頃探している。

そこで、最後にそうした目的のために役立ついくつかの統計関係の参考書をあげてみたい。本稿のレビューの対象期間に刊行されたものの中には、山口(2003)の「図解入門 よくわかる統計解析の基本と仕組み」がある。一通り統計学の授業を受けて一応の予備知識を持った学生が本書を自習すれば、知識が極めて整理されるものと思われる。櫻井・神宮(2003)の「使える統計—Excelで学ぶ実践心理統計—」も、いきなり自習するには難しいであろうが、Excelの関数や分析ツールの操作が示されているので、基本的な予備知識があれば、分析の際に参考にできる。森・田中(2003)の「なっとくする統計」は、意識調査などの調査データを分析するという視点で読者を導きつつ、推定・検定・多変量解析まで網羅している。初歩から始まる1年間4単位の授業では扱いきれないであろうが、1冊でよくまとまっていると言えるだろう。小野寺・狩野(2003a,b)の「文科系の学生のための数学入門1および2」の2冊は、確率・統計だけでなく行列、固有値、微積分などを含んでいる。大学院でテスト理論を本格的に学びはじめようとする文科系の学部生にとって手頃なテキストであろう。

各種統計解析ソフトの利用環境もさらに整ってきた。

私立文系といえども、心理学系の学部生は卒業研究でデータ分析を伴う実証的研究を行うことが多い。統計的分析に汎用されている SPSS については、学生個人が購入したい場合、石村(2003)の「CD-ROM 統計ソフト SPSS Student Version 11.0J」を手に入れることができるようになった。大学単位での一括導入ができて不自由していない大学もあるが、諸般の事情によってそれがかなわない場合でも、学生個人に購入をすすめやすくなった。データとして入力できる変数の数に制約があるものの、小規模のデータの分析ならば活用できる。この CD-ROM にテキストはついているが、その他のテキストとして、内田・牧野・長谷川・西澤(2002)の「すぐに使える SPSS によるデータ処理 Q&A」も、SPSS による分析作業の操作の一助として役立つであろう。ただ、SPSS を操作して何らかの結果を得るという作業がますます身近でたやすくなったため、そもそも適用する分析が適切ではなかったり、解釈を誤ったりという問題はなかなか減ることがない。統計学の基礎の習得には、段階を追って時間をかけた教育プログラムと学生自身の取り組みが必要である。そういった認識の広まりを今後も期待したい。

5. おわりに

大学生の学力低下が指摘されて久しい。しかし、学部教育の内容を情勢に合わせていくことになり、学部教育の充実どころか易しいレベルに定常化せざるをえなくなると、大学院教育さらには技術や理論の進歩との距離がますます広がっていく恐れがある。測定・評価の分野に関してみれば、統計学の学部教育の充実を図ることによって統計的手法の基礎知識を広く普及させるようにしていくことも大切であるが、大学院や研究所、学会などの随所でテスト理論や心理統計学の専門的・本格的な学習のプログラムが常に用意されていることも、この分野の大学院生・若手研究者が増加して発展していくためには重要なことであろう。大学院レベルの教育の維持とさらなる発展を願うとともに、関連の学会活動においても学習・啓蒙の機会がこれからも続いていくことを願う。

最後になるが、このレビューを通して、この1年、「教育心理学研究」などの学会誌へのこの分野からの掲載論文が比較的少ない印象を受けたが、関連諸学会においてテストや評価関係の研究発表は数多く見られた。来年度以降、雑誌への掲載論文が増加していくことが期待できるであろう。

引用文献

安藤雅和・野口裕之 2003 項目公開制を採る大規模試験における IRT 尺度の等化法(2)—計算アルゴリズム

- の検討— 日本心理学会第67回大会発表論文集, 359.
- Angoff, W. A. 1971 Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement*. 2nd ed. Washington, D.C.: American Council on Education. Pp. 506-600.
- Brennan, R. L. 2001 *Generalizability theory*. New York: Springer-Verlag.
- 大学入試センター 2002 国際シンポジウム 問題作成からみる大学入試 平成14年11月16~17日 独立行政法人 大学入試センター
- 原田康也 2003 IT 技術を応用した語学試験 (speaking) の新しい試み 日本テスト学会第1回大会発表論文抄録集, 67-70.
- 服部 環 2003 授業評価を行うための質問項目を作成する試み 日本教育心理学会第45回総会発表論文集, 288.
- 林 篤裕 2003 Rule Space Method とその適用例 日本行動計量学会第31回大会発表論文抄録集, 230-233.
- 林 則生 2003 「適応型テスト方式」によるインターネット受検の開発過程 日本テスト学会第1回大会発表論文抄録集, 71-72.
- 廣瀬英子 2000 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 教育心理学研究, 48, 235-246.
- 廣瀬英子 2002 プレゼンテーションにおけるスライド活用効果と内容理解度に関する実践的検討 日本教育工学雑誌, 26 (Suppl.), 51-54.
- 廣瀬英子 2003 情報教育における学生のパフォーマンスに対する相互評価の試み 2003PC カンファレンス発表論文集, 67-68.
- Holland, P. W., & Rubin, D. B. (Eds.) 1982 *Test equating*. New York: Academic Press.
- 池田 央 2003 コンピュータを利用した心理・教育テストの動向—米国を例に 情報処理学会, 情報処理, 44 (9), 940-945.
- 井上俊哉・孫 媛 2003 DIF を用いた高校学力検査データの分析 日本心理学会第67回大会発表論文集, 369.
- 犬塚美輪 2003 作文指導における文章評価活動の効果 日本教育心理学会第45回総会発表論文集, 286.
- Irvine, S. H., & Kyllonen, P. C. (Eds.) 2002 *Item generation for test development*. Mahwah, NJ: Lawrence Erlbaum Associates.
- 石村貞夫 2003 CD-ROM 統計ソフト SPSS Student Version 11.0J 東京図書
- 石岡恒憲・亀田雅之 2002 コンピュータによる日本語

- 小論文の自動採点システム 電子情報通信学会, 信学技報, **102**(491), 43-48.
- 石塚智一(研究代表者) 2002 大学入学者選抜における評価の標準化に関する研究 平成11年度～平成13年度 大学入試センター研究開発部 共同研究報告書
- 石塚智一 2003 我が国における伝統的試験と標準化された試験 日本行動計量学会第31回大会発表論文抄録集, 286-289.
- 梶井芳明 2003 児童の作文はどのように評価されるのか? (4)一学年の識別性による検討— 日本教育心理学会第45回総会発表論文集, 70.
- 柏木繁男 2003 性格テストについて— Big Five 的観点から— 日本テスト学会第1回大会発表論文抄録集, 5-8.
- 菊地賢一 2003a Microsoft Office を利用した CAT システムの開発 日本行動計量学会第31回大会発表論文抄録集, 214-215.
- 菊地賢一 2003b MS Word と Excel を利用した汎用的 CAT システムの開発 日本テスト学会第1回大会発表論文抄録集, 33-34.
- 小林典子 2003 SPOT による日本語能力の測定 日本行動計量学会第31回大会発表論文抄録集, 110-113.
- Kolen, M. J., & Brennan, R. L. 1995 *Test equating: Methods and practices*. New York: Springer-Verlag.
- 熊谷龍一 2003 語彙理解尺度の DIF 分析の試み—BIG-MG および SIBTEST を用いた DIF 項目の発見— 日本心理学会第67回大会発表論文集, 365.
- 前川久男 2003 知能テストについて—この10年における変化を中心に— 日本テスト学会第1回大会発表論文抄録集, 3-4.
- 前川眞一 2003a テストの標準化について 日本行動計量学会第31回大会発表論文抄録集, 290-293.
- 前川眞一 2003b 大規模テストにおける「等化」もしくは「共通尺度化」の重要性 日本テスト学会第1回大会発表論文抄録集, 63-64.
- Mills, C. N., Potenza, M. T., Fremer, J. J., & Ward, W. C. (Eds.) 2002 *Computer-based testing: Building the foundation for future assessments*. Mahwah, NJ: Lawrence Erlbaum Associates.
- 森 真・田中ゆかり 2003 なっとくする統計 講談社
- 森 敏昭・深谷優子・梶井芳明・村山 航・鹿毛雅治・大塚雄作・村井潤一郎 2003 あらためて教育評価を問う—「学力」を伸ばす教師および評価実践とは— 日本教育心理学会第45回総会発表論文集, S62-S63.
- 村山 航 2003 テスト形式が学習方略に与える影響 教育心理学研究, **51**, 1-12.
- 室山晴美 2002 コンピュータによる職業適性診断システムの利用と評価 教育心理学研究, **50**, 311-322.
- 室山晴美 2003 適性テストの開発と発展—職業診断システム「In ★ Sites2000」を中心に— 日本テスト学会第1回大会発表論文抄録集, 9-12.
- 村上 隆(研究代表者) 2003a 我が国の公的試験における得点等化の導入に向けた心理・教育測定的研究 平成12年度～平成14年度 科学研究費補助金(課題番号 12800015) 研究成果報告書
- 村上 隆 2003b 何が日本の入試得点の標準化を阻んでいるのか 日本行動計量学会第31回大会発表論文抄録集, 298-301.
- 夏堀 睦 2002 学校内での物語評価についての一般的信念の検討 教育心理学研究, **50**, 334-344.
- 夏堀 睦 2003 教師の物語創作評価の分析 日本教育心理学会第45回総会発表論文集, 77.
- 日本語教育学会(編) 2003 平成13年度日本語能力試験分析評価に関する報告書 国際交流基金・日本国際教育協会
- 日本教育心理学会(編) 2003 教育心理学ハンドブック 有斐閣
- 仁田善雄・奈良信雄・吉田素文・寺尾俊哉・石田達樹・福島統・齋藤宣彦・福田康一郎・高久史磨・麻生武志 2003 第2回医学系 CBT 共用試験トライアルの統計解析 日本テスト学会第1回大会発表論文抄録集, 25-26.
- 野口裕之・安藤雅和 2003 項目公開制を採る大規模試験における IRT 尺度の等化法(1)—等化係数の推定法— 日本心理学会第67回大会発表論文集, 358.
- 野口裕之・渡辺直登 2003 項目反応理論によるパーソナリティ／態度テストの開発と検討—産業・組織心理学領域におけるフォアフロント— 日本心理学会第67回大会発表論文集, S77.
- 小方博之 2003 ハイパーグラフによる数式問題答案の評価 日本テスト学会第1回大会発表論文抄録集, 31-32.
- 小野寺孝義・狩野 裕 2003a 文科系の学生のための数学入門1 ナカニシヤ出版
- 小野寺孝義・狩野 裕 2003b 文科系の学生のための数学入門2 ナカニシヤ出版
- 大澤公一 2003 多値型項目反応モデルによる評定者特性の推定—多相モデルと階層モデルの比較を中心に— 日本テスト学会第1回大会発表論文抄録集, 35-38.
- 大塚雄作 1994 理論と実践を結ぶ測定・評価に向けて 教育心理学年報, **33**, 102-111.
- 尾崎幸謙・豊田秀樹 2003 3段階反応型採点による一

- 対比較 IRT モデルの提案 日本行動計量学会第31回大会発表論文抄録集, 222-223.
- Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (Eds.) 2002 *Practical considerations in computer-based testing*. New York: Springer-Verlag.
- Petersen, N. S., Kolen, M. J., & Hoover, H. D. 1989 In R. L. Linn (Ed.), *Educational measurement*. 3rd ed. New York: American Council on Education and Macmillan Publishing Company. Pp.221-262. (池田央・藤田恵璽・柳井晴夫・繁榊算男(訳編) 1992 教育測定学 第3版 上下 C.S.L.教育学習評価研究所)
- 三枝令子 2003 日本語 Can-do-statements の開発 日本行動計量学会第31回大会発表論文抄録集, 102-105.
- 佐伯圭一郎・西山悦子・高木廣文・柳井晴夫・道場信孝・日野原重明 2003 生活習慣テストについて—LPC 式生活習慣検査の開発— 日本テスト学会第1回大会発表論文抄録集, 13-16.
- 斉田智里 2003 高校入学時の英語能力値の年次推移—項目応答理論を用いた県規模英語学力テストの共通尺度化— STEP Bulletin, 15, 12-24.
- 斉田智里・服部 環 2003a 高校英語テストのための項目プールと難易度別テストの作成 第45回日本教育心理学会総会, 411.
- 斉田智里・服部 環 2003b 県規模英語学力テストの垂直的等化—項目応答理論の大規模英語学力テストへの適用と学年別英語能力値の変化— 日本心理学会第67回大会発表論文集, 1205.
- 櫻井広幸・神宮英夫 2003 使える統計—Excel で学ぶ実践心理統計— ナカニシヤ出版
- 佐藤達夫 2003 医学教育における共用試験—その意義と試験問題の共有化 日本テスト学会第1回大会発表論文抄録集, 65-66.
- Shermis, M.D., & Burstein, J.C. (Eds.) 2003 *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.
- 椎名久美子 2003 標準化試験のための作業段階—SAT, TOEFL の事例— 日本行動計量学会第31回大会発表論文抄録集, 294-297.
- 荘島宏二郎 2003a 項目反応理論 (IRT) における等化 日本心理学会第67回大会発表論文集, S19.
- 荘島宏二郎 2003b 実験計画のための項目反応モデル 心理学研究, 73, 506-511.
- 荘島宏二郎・豊田秀樹 2002 項目反応理論における Cronbach の α 係数の推定 心理学研究, 73, 227-233.
- 荘島宏二郎・吉村 幸・中畝菜穂子 2003 項目反応理論から見た大学入試センター試験「英語」の特徴 日本テスト学会第1回大会発表論文抄録集, 51-54.
- Thissen, D., & Wainer, H. (Eds.) 2001 *Test scoring*. Mahwah, NJ: Lawrence Erlbaum Associates.
- 内田 治・牧野泰江・長谷川博康・西澤英子 2002 すぐに使える SPSS によるデータ処理 Q & A 東京図書
- 内田照久 2002 大学入学者選抜資料の評価の標準化に対する意識調査 石塚智一(研究代表者) 大学入学者選抜における評価の標準化に関する研究 平成11年～13年度 大学入試センター研究開発部 共同研究報告書 19-31.
- van der Linden, W. J., & Glas, C. A. W. (Eds.) 2000 *Computerized adaptive testing: Theory and practice*. Dordrecht, The Netherlands: Kluwer Academic Publishers.
- von Davier, A. A., Thayer, D. T., & Holland, P. W. 2003 *The kernel method of test equating*. New York: Springer-Verlag.
- 和田晃子 2003 オーストラリアにおける日本語教育測定 日本行動計量学会第31回大会発表論文抄録集, 114-117.
- 山口和範 2003 図解入門 よくわかる統計解析の基本と仕組み 秀和システム
- 山森光陽 2002 一般化可能性理論を用いた観点別評価の方法論の検討 STEP Bulletin, 14, 62-71.
- 山森光陽 2003 中学校英語科の観点別学習状況の評価における関心・意欲・態度の評価の検討—多変量一般化可能性理論を用いて— 教育心理学研究, 51, 195-204.
- 柳井晴夫 2003 「心理テストの過去・現在・未来」のねらい 日本テスト学会第1回大会発表論文抄録集, 1-2.
- 吉村 幸・中畝菜穂子・荘島宏二郎・石岡恒憲 2003 大学入試センター試験統計情報データベース作成の試み 日本テスト学会第1回大会発表論文抄録集, 43-46.

謝 辞

本稿の執筆にあたり、池田 央 立教大学名誉教授には、関連資料について多くの示唆と助言をいただきました。ここに感謝いたします。