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―

近年コンピュータを利用したテスト形式（CBTと呼ばれる）の急速な実用化が進んでいる．ここでは先端を行く米国を主たる事例に，進展の歴史的経緯と現況を整理し展望することにした．新しいタイプのテストが持つ出題提示方法，解答方法の特徴，またその狙いに触れ，従来のペーパーテストによる試験形態との違い，またそれを実現可能にしている現代の測定理論と具体化するのに必要な技術的革新についてのさまざまな試みを説明した．さらに，それらを日本で展開する場合に考えられる難しさや必要な条件についても触れた．

池田　央
立教大学／（株）教育測定研究所
hikeda@rikkyo.ac.jp

テスト技術の革新

IT技術の進歩とともに，それを応用したテスト方法の先進諸外国（特に米国）における発達は看過できない状況にある．それは心理・教育測定の理論的基盤の発達に支えられて，入学試験や各種検定試験の大規模公式試験において顕著であり，また，e-Learningなどの学習システム設計においてもその技術は多くの発展可能性を持っている．

それは，テスト問題を作成し実施する側にも，また受験する側にも，従来から慣れ親しんできたペーパーテストからくる常識を超えて，ある意味で，テスト方法あるいは能力測定法に対するイメージをすっかり変えてしまうだけのパラダイム変換を要求するものと捉えることができるからである．

本稿では，はじめにここに至る心理・教育測定学の発達の経緯を簡単に振り返った上で，多くの人に今日のCBTの現状と将来に向けての動向を知ってもらうための解説をすることにした．ただそれは，2000年当時の状況から考えたものであり，将来どのような発展の経緯を辿るか，残るものもあり消えるものもありで，今後の予測が難しい部分も多いことをお断りしておきたい．

歴史的背景

心理・教育測定学が誕生したほぼ20世紀初頭から今日に至るまで，テスト技術という観点から分類するとコンピュータ登場前の半世紀とその後の半世紀の2つに分けて考えることが有用であろう．

前半世紀の中で考えると，テスト史で特記すべきことが2つある．1つは1910年代，いままでの論述式試験の採点結果がきわめて主観的であり恣意的であるという批判研究があり，おりしも第1次大戦時に登場した選択式客観テスト形式が，大量受験者の集団処理の要求に応え得たということが大きかった．そしてもう1つは1920年代に，難易度の異なるテスト得点を比較可能にするために，集団の平均値と標準偏差を尺度の原点と単位の基準にとる標準得点方式の提案がなされたということであろう．これら2つは今日多肢選択式問題（Multiple-Choice Question：MCQ）と呼ばれる質問形式と標準得点またはZ-得点（日本では通称偏差値）と呼ばれる得点表示形式の原型であり，その考えは今日でも広く大量試験で使われていることは周知のとおりである．

上述のテスト形式が，個の情報を集積処理する技術としてきわめて優れていたことは次の例を見ても分かる．新聞，雑誌，ラジオ，テレビといったマスメディアは，

	テスト名	目的用途	Web サイト（URL）
1	TOEFL	外国人留学生英語テスト	http://www.toefl.org/toeflcbt/cbtindex.html
2	同上	日本代表部	http://www.cieej.or.jp/Toefl/cbt/index.html
3	TSE	口頭英語コミュニケーション能力	http://www.toefl.org/tse/samples.html
4	GRE	大学院入学試験	http://www.gre.org
5	GMAT	ビジネススクール入学試験	http://www.gmat.org
6	ASVAB	軍隊適性試験	http://www.goarmy.com/util/asvab.htm
7	ARE	建築家登録試験	http://www.ncarb.org/are/tutorial2.html
8	USMLE	医師免許試験	http://www.usmle.org/step3/s3sam.htm
9	AP	大学進学上級問題	http://www.collegeboard.com/ap/students/apcd/html/apintro.html
10	IEA	Intelligent Essay Assessor	http://www.knowledge-technologies.com
11	同例	論文自動採点システム 1	http://www.k-a-t.com/HRW12Demo/HRW12.html
12	Intellimetric	論文自動採点システム 2	http://www.intellimetric.com/demosite/demo.html

表-1　各種CBT問題例の見られるWebサイト

1 つの情報発信元から多数の受信者へ情報を伝達可能にする高度の技術を発展させた．それに引き換え，散在する個別の情報を 1 つに集める技術は，いままできわめて貧弱であった．教育において最も大切なことは，学習者個別の学習状況を短期間にフォーマルなかたちで集め，それを次の学習に個別に反映させる技術である．それを集個技術と呼びたいが，それが可能なかたちは，いままでのところ客観テスト形式やアンケート調査形式くらいしかない．多肢選択法による質問や，5 段階評定法による質問は，個別の多様な情報を，対象となるすべての人（受験者）から比較可能な一定のフォームで集めることができる恐らく唯一の方法である．

1935 年すでにIBMは最初の自動採点機を試作したが，第 2 次大戦後のコンピュータと光学式読み取り装置（OMR）の登場は，大量データの高速採点処理と，得点の比較可能なかたちへの変換処理にその威力を発揮し，今日に至っている．

しかし，集団の平均値と標準偏差を使った線形変換を基礎に相対評価する通称偏差値方式は，厳密には同一集団にしか適用できないものである．1950 年代になると，集団の特性とテスト問題の特性を切り離して独立に考えようとする項目応答（反応）理論（Item Response Theory：IRT）が登場した．それは，問題項目の特性と受験者の能力とを非線形の確率関数で結び付けようとするもので，計算には一層の複雑性を要する．しかしそのおかげで，実施年の異なるテスト結果や学年の異なるテスト得点を共通の尺度上で比較可能に変換する道も開けた（これをテスト得点の等化問題という）．

今日のTOEFLスコアもそうして変換された共通尺度が用いられ，毎回異なる問題が使用されても結果は共通尺度上に表示される（項目応答理論などの解説については池田（1994）[1]，またWebサイトではhttp://ericae.net/irt/などが参考になる）．しかし，1960 ～ 70 年代は，理論的研究は進められたものの大量実用化には至らず，コンピュータの能力がそれ以上に向上するのを待たなければならなかった．

1980 年代に入り，パーソナルコンピュータが普及し，受験者 1 人 1 人がパソコンを通じて，リアルタイムで問題提示と即時応答ができる時代を迎えて，真剣にCBT化への取り組みが考えられるようになった．そして約 10 年の実験的試行の時代を経て，90 年代に入ると，本格的大規模コンピュータ試験時代を迎えることになった．今ではCBTによる大規模試験の受験者は公式的なものだけでも，世界で年間数百万人に上るといわれている．テスト形式も単に選択式テスト問題から，より本格的なシミュレーションテストやマルチメディアテスト問題へと開発が続けられている（そうした新しい試みの紹介としてやや古くなったが，Drasgow & Olson-Buchanan（1999）がある[2]）．

新しいテスト（CBT）の特徴

表-1 には米国で比較的早くから実施されているいくつかの大規模CBT試験の略称とその内容の理解できるWebサイトをあげておいた．これらを手がかりに，まずCBTの実施形態がどのようなものであるかを見ていこう．米国の代表的試験機関Educational Testing Service

（ETS）は米国内はもとより世界各所に試験センターを設けているが，志願者は事前予約により，基本的にオープンされている日ならいつでも都合のよい日を選んで，受験できる態勢がとられている．ちなみに日本人がよく利用するTOEFLは東京2会場のほか，横浜，大阪会場のいずれかで受験することができる（表-1（2））．

私が訪れたことのあるロサンゼルス郊外の試験センターを例にとると（ほかでもほぼ同様である），我々が持つ従来のテスト会場に対するイメージとは異なり，多くは1会場あたり，十数人収容程度の大きさでそれほど広いものではない．ブースに囲まれた机上のパソコンに向かって，各受験者はむしろ落ち着いた雰囲気で受験することができる．会場の外の待合室では，次の受験者が静かに待っており，それは病院の診察室のイメージに近い．サーバ管理からいってもその程度の大きさが，いちばん扱いやすいようである．

試験室内に入るときは，受験者はIDチェックを受け，荷物をロッカーに預け，指定された机に座る．試験開始前には解答方法に慣れるためのいくばくかの練習問題が用意され，音声ボリュームの調節をはじめ，あらかじめ申請が認められれば，障害者のためのトラックボール使用や文字画面のズームアップとか色彩調節もできるようになっている．こうした機能は障害者に優しいテストだとして歓迎されている．頭上に設置された鏡を利用して，受験者の様子が隣室の監督に見えるようになっている．解答には机上に置かれたメモ用紙を使ってよいが，退室時には残して出なくてはならない．

問題項目は日によって，あるいは人によって変えられるが，このことは一斉テストで全員が同一問題セットを解かなければならないといういままでの制約条件から開放される．また，同じ日同じ時間に，大量のパソコンを備えた会場を用意する必要もない．したがって年に僅かの特定日だけに大勢の受験者が集まり，あとの日は会場のパソコンや採点機械が遊んでいるということは起こらない．常時誰かが受験しておればよく，またそれは同じ試験の受験者である必要もない．ある机では，大学院に入るための受験者がいる一方で，となりの机では別の資格試験の受験者が受けているということもあり得る．こうして年間常時，ほぼ一定の受験者が回転していけば，試験センターの運営も楽になるというものである（実際はどうしてもある時期に受験者が集中したりしてうまくいかないことも多いようであるが）．

このように，出される問題が受験者によって異なるのに，どうして受験者の能力を公平に査定することができるのだろうか．その問題を解決したのが，先にあげた能力と問題難易度を切り離して独立に推定可能にした項目応答理論（IRT）である．全員が同じ問題を与えられ，正答できた量で人の能力の高低が定義されるという従来のテスト方式と違って，受験者の能力を見ながらそれに合わせて出題を決めるという適応型テスト（Computerized Adaptive Testing：CAT）では，正答数の最終結果はほぼ全員が50%近くになり，そこでは差がつかない．能力差が決まるのは正答した数の大小ではなく，どの問題が正答できたかが重要なのである．これはテスト得点に対する考え方のパラダイム変換であり，受験生にもその方式に慣れるまでは，多少の違和感があるようである．**図-1**は適応型テストによる能力推定のプロセスを概念化して描いたものである．

しかし，このようなことが可能になるためには，その背後に各問題の価値にかかわる項目特性値の分かっている既知の大量問題項目群（問題項目バンクと呼ばれる）が存在し，たえずそのような問題項目を補充維持しておくシステムが必要である．TOEFLでも何題かの予備問題と称される採点されない問題が含まれるのはその準備のためである．

試験の得点は，正式には後日受験者に郵送通知されるが，希望すれば終了後直ちに知ることができる．いままでのように何日も結果が報告されず，受験者がいらいらしていたことはなくなり，即結果が分かるということが，受験者にとっていちばん好評のようである．それこそコンピュータ試験ならではのことである．Wainer（2000）[3]をみると，こうした新しいテストの姿がうまく描かれている．

いまは，受験者本人の同定と，セキュリティ上の問題から，受験は試験センターで行われているが，この問題が解決すれば，Web上での試験実施もやがて日常的なこととなっていくであろう．そうなれば，自宅や自分のオフィスからの受験も可能になり，文字通り時と場所を問わないOn Line Testingが実現する．そしてその技術は，何も試験場面だけでなく，インタラクティブで即時フィードバックを必要とするコンピュータを使った遠隔授業やe-Learning教材の中にも活かされて，教育効果を高めることに役立つであろう．

CBT化への移行と方略

もちろんそうした仕組みが一夜にしてでき上がるわけ

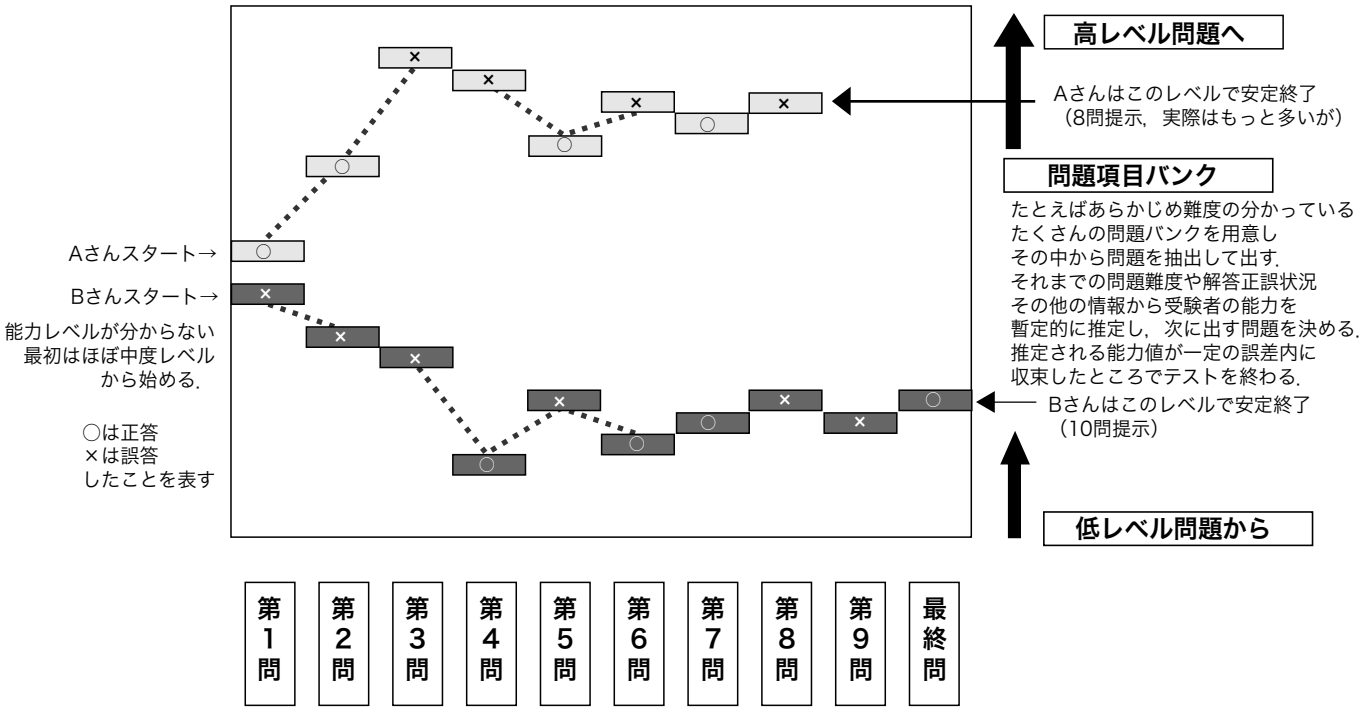


図-1　適応型テストの仕組み

ではない．初期の研究段階から現在に至るまでのCBTの発展経過を見ると，当然ながらいくつかの段階に分かれるようである．適応型テストの実用化までの経緯をまとめた報告書としてSands, Waters, & McBride（1997）は参考になる[4]．

CBTの開発は出題形式の変遷と解答方法の多様化と2つの面から考えるとよい．

第1段階：ペーパーテストのコンピュータ化

それは現在あるマークシート式テスト問題をコンピュータディスプレイ上に表示し，与えられた選択肢をマウスでクリックするか，数字をキーインして答える形式のものである．当然ながら，いちばん手っ取り早く，誰しもが考えることで，このかたちを取っているコンピュータ試験は多い．

しかし，それはペーパーテストでできることを，コンピュータに移し変えただけで，採点結果がすぐに出ること以外に，コンピュータを利用するメリットはあまり大きくない．マークシート式に比べて経費がかかること，設備を必要とすることを考えるとかえってマイナスの方が大きいともいえる．ただ，同じ試験でも，設備のあるところではコンピュータで行い，そうでないところではペーパーテストで実施するという併用制がしばらく続くことがあるかもしれない．そのときはコンピュータテストとペーパーテストのどちらを受けても差がないという等価性が問題となり，コンピュータ用テスト問題とペーパー用テスト問題を変えるわけにいかなくなる．それを続けることはいつまでたってもコンピュータ特有のテスト問題に切り替えることができないというジレンマに陥り，注意しなければならない．

第2段階：出題方法の多様化－マルチメディアの活用

テストのコンピュータ利用は，それによって従来のペーパーテストでは測り得なかった能力を測ることができるようになるということである．そのためには出題方法と解答方法を工夫しなければならない．

解答方法を選択式にすることは，あとの採点を考えると，集個技術としてはきわめて優れた方法である．その特徴を残しながら，出題を多様化するという方法が次に考えられることである．

問題の提示には，従来のような主として印刷された文字，あるいは高々図形や写真などを貼り付けて利用する

形式	用途（利点）
Drag & Drop	空所補充，移動
Cut & Paste	単語，文節の入れ替え，並べ替え
Hot Spot	地図上小エリアの選定
Radio Button	多肢選択（重複入力の回避）
Check Box	複数選択の許容
List Box	大量選択肢
Combo Box	大量選択肢，文字入力ミス回避
Slider	微小連続数値などの入力
Keyboard	文字，数字，記号入力
Touch Panel	幼児，年少者，障害者向け

表-2　選択式問題の主な解答入力例

以外に，音声，画像（写真やグラフ），アニメ，動画などを利用して，提示することが考えられる．これをマルチメディアテストと呼ぶことにしよう．特に語学の聴解力試験（Listening Test）や科学実験，音楽，あるいは地図や芸術品を題材にする地理・歴史の試験などに向いている．

しかし，それには問題作成に経費がかかり，材料をモジュール化し，必要とする人がそれを共有化できるようにアイテムバンク化（実際にはこれが難しい）のシステム作りが進行しないと普及は難しかろう．そういう意味でのインフラ整備が重要である．また試験の実施には受験者側のコンピュータ上に画像提示やヘッドホンなどの設備が必要になってくることは当然である．CBTの実用化にあたって生ずるいろいろな問題点や工夫については Parshallほか（2002）[5] が参考になる．

第 3 段階：解答方法の多様化－多肢選択式を超えて

出題方法の多様化と並行して，コンピュータの入力機能を利用して，解答方法の多様化も考えられてきた．しかし受験者にそのために特殊な技能や過度の練習を必要とするものは好ましくない．ペーパーテストの基本形は受験者が鉛筆で解答を記入することが前提であったが，それに代わるものはCBTではマウスまたはキーボード入力である．

しかし，自由な記述式テストは採点の自動化を考えると，最後まで残る課題であり，当面の課題は，まったく自由な解答を許す前に，あり得る多様な解答をいかに標準化し，規格化して有限個の解答パタンに絞り込むかを考えることであろう．選択式問題の利点を残しながら，解答の自由度をもう少し高めたものとして，**表-2** のような方式が考えられる．

与えられたオプションから 1 つを選ぶという多肢選択式問題（MCQ）は，能力の一面しか見ていない，皮相的な力しか見ていない，と古くから批判の対象となってきた．また最近の教育目標としても，求められるのは「何を知っているか」（What you know?）だけでなく「何ができるか？」（What you can do?）の方である，すなわち知識偏重から知識＋技能つまりパフォーマンス重視への傾斜傾向がある．

こうした要求に応えるために，テストに求められる解答形式も，与えられたオプションから解答を選ぶものだけでなく，解答者自身が解答を構築していくタイプの問題が増えてきている．それを解答構築型のテスト問題（Constructed-Response Question：CRQ）という．コンピュータの持つ多様な機能は，その構築を助ける．そのために，計算，作図，作画のツールボックスを利用するもの，リンクすれば表計算ソフト，統計表，辞書検索などが利用できるようにするものなどがある．表-1 の中身を探していけばそうした新しいタイプの試験問題例を見ることができる．

第 4 段階：シミュレーションテスト，VRテストへ

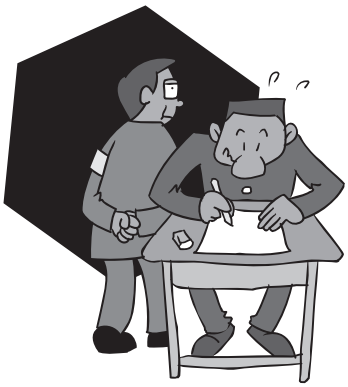
テストで知りたい究極目標が，受験者が実際に予想される仕事場面で，能力をどれだけ発揮できるか見たいということであれば，できるだけ本物の仕事そのものに近い見本を実際にやらしてみればよい．これが真生テスト（Authentic Test）である．英会話能力であれば実際に話しているのを見れば分かるし，医者であれば，実際に患者の病状を確かめ，的確な診断ができて，適切な処置ができるかを見ればよい．また，仕事に必要な優れた文書を構築できる能力があるか確かめたければ，実際に課題を与えて文章を書かせてみればよい．ペーパーテスト，特に選択式のテスト問題では，どうしてもそれは間接的な測定になり，見たい対象そのものの測定になっていない．

実物そのものにはなり得ないが，コンピュータ上の仮想現実（Virtual Reality：VR）は，それに近い場面を作り出す可能性を持っている．また最近では，仕事自体がほとんどコンピュータ上の画面を操作することによってなされる場面が増えている．建築家の登録試験（ARE）や医師免許の試験（USMLE）の一部であるPatient Management Problem（PMP）などはそうしたシミュレーションテストの例である（表-1（7），（8））．

第 5 段階：採点技術の開発－これからの課題

こうして，CBT化によって，出題形式と解答方法の多様化は急速に進み，新しい解答構築型のテストやシミュレーションテストも次々と生まれてくるが，やはり残された難題は，その解答結果をいかに得点化するかという採点の基礎理論である．選択式問題のように，すでにコード化された解答を処理するには項目応答理論（IRT）も威力を発揮するが，その前に自由な解答をいかにコード化して，意味あるかたちの数値パタンに変換するかという計量心理学的理論が必要となってくる．そうでなければいままでのように人間の主観的判断で評定する作業が残る．人間の判断過程に代わって，どこまでコンピュータの自動化が可能か，現在さまざまな試みがなされているが，まだ確固たる理論形成には至っていない．最近では英語の論文テスト（Essay Test）の自動採点も実用化の段階に達している（表-1（10），（11），（12））（Shermis and Burstein（2002）[6]）．ただ，それらの方式の技術的仕様については，次第に知的所有権をめぐって企業秘密化していく傾向があり，なかなか細部にわたって知ることが難しくなってきている．

我が国で一挙にその段階にいくことは難しいとしても，過渡期として，論文テストの手採点部分などの一部工程を，コンピュータの力を借りて，合理化することがもっと考えられてもよいのではなかろうか．たとえばワープロで入力解答された記述式答案をあらかじめ語数順に並べ替えて，採点者に出力提示させれば，キャリーオーバー効果と呼ばれる前出答案の影響を緩和できるのではないかとか，複数者による採点の場合は他の評定者の採点状況が同時に提示されて参考にできるとか（それがよいか悪いかは別として），自分が今までにつけてきた採点状況が何人分か採点するごとに経過報告され，マニュアルで指定された目安の得点配分と大幅にずれていないかチェックできるなど，まだ採点工程システムとして工夫される余地は十分残されているように思う．



我が国の課題

我が国でもこうした米国の例などに刺激されて，このところ急速にコンピュータテストへの関心が高まってきた．米国が辿ってきた開発，実用化の道を追いかけるのは容易なことではないが，それを抜きにするわけにもいくまい．適応型テストの実施には，問題項目バンクの構築も必要である．高度の技術を駆使し，想定される危険を克服して，安全な実施が可能になるまでには，長期の実験期間が必要であろう．拙速はかえって批判を呼び，着実な研究開発が阻害される恐れも大きい．

むしろこのテスト技術は我が国ではe-Learningの枠組みの中で，形成的評価のための学習ナビゲータの役割が期待されるのではあるまいか．個人ごとの学習経過をモニタし，次への学習プログラムをガイドしていくためには，用意された問題項目への解答パタンから全体像（全体値）の中での学習者の位置付け（現在値）をはっきりさせ，そこからその人の次なる学習ターゲット（目標値）を設定し，それに近づけるための学習課題を用意するというカリキュラムと一体化された継続測定，すなわちダイナミックテスティングと呼ばれる新しいテストの方向性が，これから目指すコンピュータテストの道ではないかと思われる．

それにしても，我が国ではまだテストの作成，分析，採点，利用法の各技術を専門的に研究する研究者がきわめて少ない．また項目特性値の分かった利用できるテスト問題もきわめて少ない．こうしたインフラ作りも，我が国が早急に充足を図らねばならない課題である．

参考文献

1）池田　央：現代テスト理論，朝倉書店，東京（1994）．
2）Drasgow, F. and Olson-Buchanan, J. B.(Eds.): Innovations in Computerized Assessment, Lawrence Erlbaum Associates, New Jersey (1999).
3）Wainer, H.(Ed.): Computerized Adaptive Testing: A Primer (2nd. Ed.), Lawrence Erlbaum Associates, New Jersey (2000).
4）Sands, W.A., Waters, B.K. and McBride, J.R.(Eds.): Computerized Adaptive Testing, -From Inquiry to Operation, American Psychological Association, Washington, D.C.(1997).
5）Parshall, C.G., Spray, J.A., Kalohn, J.C. and Davey, T.: Practical Considerations in Computer-Based Testing, Springer, New York (2002).
6）Shermis, M.D. and Burstein, J.(Eds.): Automated Essay Scoring, Lawrence Erlbaum Associates, New Jersey (2002).

（平成 15 年 6 月 26 日受付）