テストの歴史的変遷と
コンピュータ適応型テストの意義

小　山　由紀江

This paper first describes the history of testing, especially that of language testing, in order to provide a background to Computer Based Testing and Computerized Adaptive Testing. Secondly, the paper introduces an implementation of CAT, whose content is English for general science and technology, along with a description of its construction and results. The results indicate that the CAT is more efficient than TOEIC and final examinations at NIT in terms of time consumed and the number of items answered. The result has relatively high correlation coefficients both with TOEIC and the final examination, which suggests that the CAT could be reliable and valid substitute for other tests. However, the termination condition and the difficulty level of items should be carefully considered in order to further increase to the validity and reliability of CAT. In addition, the quality and the quantity of the item-bank is crucial.

1．始めに

コンピュータ適応型テスト（Computerized Adaptive Testing；以下CAT）はコンピュータが広範に使用されるようになった1990年代からアメリカを中心に実施され、時間や人手をかけずに正確に受験者の能力を測定するテストとして言語テストの分野でも注目を集めてきた。しかし、CATが現実に使用されるようになるには、まず1980年代にComputer Based Testing（CBT）が開発され欧米諸国を中心に広範に使用されることによって、研究者によってそのメリットが明らかにされてきたという経緯がある。

本論文では言語テストの歴史的な流れを俯瞰した後、CBT に始まったCAT を言語テストとして用いることの意義と問題点について論じる。さらに、筆者が行った小規模なCAT 実験の結果と考察を加えることによって、CAT の今後の課題と可能性を提示するものである。

2．テストの分類

2．1　伝統的測定法と心理測定法

テスト理論は、様々な観点から分類されるが、テストの歴史的推移を考える場合 1）pre-psychometric measurement（心理測定法以前）と 2）psychometric measurement（心理測定法）という分類から始めるのが妥当であろう。心理測定法は、それ以前の「主観的」伝統的採点法に対する客観的な統計的採点法として、1930年代にアメリカを中心に広まった。1890年代までヨーロッパで行われていた大学入学試験や、20世紀初頭まで行われていた中国の「科挙」のような試験は出題の形式、内容、採点方法において経済学者で統計学にも優れた業績を残した Edgeworth（1888）が "unavoidable uncertainty" と呼んだところの何らかの uncertainty（不確実性）を含むテストであった。一方同じく1890年ごろから Thorndike らによって子供の精神的な発達を測定し発達の遅れを診断するテストが開発され1912年には Thorndike の弟子が作文を採点するスケールを作成した。このような教育界の動きが1920年代の「新しい」言語テスト開発に繋がっていった。Spolsky（1995）はこの一連の動向を以下のように巧みに表現している。

> Modern objective language testing evolved at a time when the new-type tests were starting to offer a tempting solution to the statistical challenge to examinations implicit in Edgeworth's assertion of the 'unavoidable uncertainty' of measurement. (p.53)

Spolsky（1995）はさらに「この統計的な測定は『何を測っているか』はともかくも測定の一貫性だけは確保し、その意味で信頼性の問題を解決しているようにあらゆる点から見えた。」と述べている。しかし続けて、エッセイの採点に関する疑問などこれに反対する勢力についても詳述しており、「客

観的」「統計的」測定がテストの抱える全ての問題を解決したわけではない点にも目を向けておかなくてはならないことを指摘している。

以上述べたように、この「新しい」テストが心理的精神的発達を測定するテストの開発に始まったことから、テスト結果の統計的分析と解釈を伴う測定法は psychometric measurement（心理測定法）と呼ばれ、歴史的には言語テストの領域にもこの呼び方が適用されることになった。

2．2　古典的テスト理論と項目応答理論

心理測定法と呼ばれるテスト法は1920年代には「新しい」テスト法であったが、その後このテスト分析や解釈の理論は後述されるような様々な限界を指摘されるようになった。そして、現在では項目応答理論（Item Response Theory；IRT）との対比で古典的テスト理論（Classical Testing Theory）と呼ばれるようになっている。IRT は、1950年代になってこの古典的テスト理論の問題点を解決する理論として提案されたテスト理論であるが、また昨今はさらに IRT の問題点を克服する理論として荘島（2009）によって「自己組織化マップや生成トポグラフィックマッピングのメカニズムを利用した統計モデル」としてニューラルテスト理論（Neural Testing Theory）も提言されている。これは能力を測定する時に連続尺度ではなく順序尺度を仮定し、段階評価を想定したテスト理論であるため、能力と得点の対応関係を説明しやすいという利点がある。

さて、上述した古典的テスト理論の限界はそれが集団準拠測定法（Norm-referenced measurement）に基づいている点にある。即ち、平均点、標準偏差等の教育界でよく耳にしてきた得点に関わる情報は受験者集団全体の能力に依拠する。つまり能力の高い受験者集団（A）が受験した場合は、能力の低い受験者集団（B）が受験した場合に比べ、平均点や最高点が当然ながら高くなる。従って、あるPという同じ受験者が（A）と（B）という二つの異なる受験者集団の中で受験をした場合、Pの得点自体は変わらなくても偏差値や順位は異なる結果となる。これは受験者集団の中で能力が相対的に評価されることが原因である。相対的な位置関係は分かっても能力そのもの（絶対的評価）は明らかにならない。さらに言えば、大友（1996）が言うように「得点」が持つ意味自体にも疑問が生じる。テストの得点は長さや重さの単位であるメートルやキログラムのように、1点が絶対的単位として同じ

意味を持っているのかという点も、試験の難易度によって点数が左右されることを考えれば極めて不明確であることが分かる。ある易しい問題に正解した場合の得点も、難しい問題に正解した場合の得点も、同じ1点であるわけだが、これらの1点の意味は異なるはずである。以上述べたように、古典的テスト理論は様々な統計的な分析を行うが、上記のような曖昧さを生得的に持つ得点というものを根拠にした統計であることは否めない。ここに古典的テスト理論の限界が指摘されることになる。

これに対しLord（1950）が最初に提言した項目応答理論（Item Response Theory：IRT）はそれぞれの項目（問題）に難易度パラメータを付加することによって、受験者の絶対的な能力を測定することができる方法である。大友（1996）のまとめたIRTの利点は以下のようなものである。

1）どのような異なったテストを用いても共通の尺度上で能力測定が可能（Test-free person measurement）：被験者の能力推定値は被験者に実施された特定のテスト項目と切り離して独立に求めることができる。
2）どんな受験者集団に実施しても、共通の項目特性に関する値を求めることが可能（Sample-free item calibration）：困難度パラメータ等の項目特性は、受験者集団とは独立して求めることができる。
3）能力ごとに分かる測定の精度（Multiple reliability estimation）：項目情報関数（item information function）によって各受験者の測定精度が示される。（p17－20）

以上まとめられたように、IRTは古典的テスト理論の限界を打ち破り、受験者集団と関係なく個々の項目の独立した困難度パラメータを求めることができる革新的な理論と言うことができるだろう。

2．3　Computer Based Testing と Computerized Adaptive Testing

Computer Based Testing（CBT）とComputerized Adaptive Testing（コンピュータ適応型テスト：CAT）は、本質的には2．2で述べられたIRTの登場と、コンピュータの発達により可能となったわけだが、1970年代には既にテストの専門家たちによって多くのCATの研究が行われていた。このCBTとCATの歴史的推移についてはChalhoub-Deville（2001）に詳述されている。Chalhoub-Deville（2001）によると言語テストの分野ではCBTとCATの登場はやや遅れ、1985年のLTRC（Language Testing Research

& Colloquium）まで待たなくてはならなかった。このLTRCの発表原稿は後に *Technology and Language Testing*（Stansfield, 1986）として出版されたが、ここにCBTとCATに関する論文が多く掲載されていたことが一つの契機となって、その後多くの教育機関でCBTやCATの実践と研究が行われるようになった。そしてその集大成とも言うべき著作がDunkel（1991）の *Computer assisted language learning and testing: Research issues and practice.* でありここではさらに多くのCATの実践とそれらに対する考察が展開されている。

CBTの積極的意味は、コンピュータでしか実現し得ないテスト方法にあるわけだが、初期のCBTを見ると中にはpencil & paper式のテストをコンピュータに置き換えただけのものもあった。しかし、それでも以前のテストとは異なるCBTの基本的なメリットがありそれはBrown（1997）によって次のように説明されている。

1）Computer-assisted language tests can be individually administered, even on a walk-in basis. Thus group-administered tests and all of the organizational constraints that they impose will no longer be necessary.

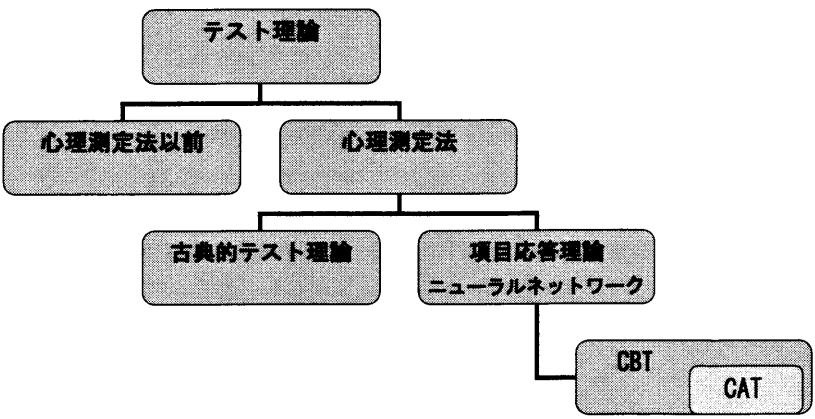
2）Traditional time limits are not necessary. Students can be given as much time as they need to finish a given test because no human proctor needs to wait around for them to finish the test.

即ち、個々人が都合の良い時に受けることができるという点と、試験監督が必要ないため時間的制約に縛られないという点である。後者については試験の性格によって時間制限の必要なテストもあるが、コンピュータを使えばそれも問題ごとに制御することが可能であり、回答時間などの受験者の回答行動もコンピュータ操作の履歴等から分析することが可能である。

一般的にはCBTにはさらに次のようなメリットがある。即ち、一度適切なテストが出来上がれば、コンピュータ等の環境が整っている限り、どこで何時、誰が（何人）受けようと、瞬時に結果が出され、必要であれば受験者にその結果をすぐ告知することができるという点である。またグラフィカルな出題内容や、ビデオや音声を使用した問題などコンピュータでしか実現しないテストもある。さらに受験者の解答行動もすべて履歴データと蓄積されるため、これを分析することによって試験の問題点を明らかにし

たり、受験者の傾向を把握したりすることもできる。CBT としてこれまで広範に実施された例としてはアメリカにおける GRE（the Graduate Record Examination), Medical College Admission Test, TOEFL iBT 等があるが、とりわけ言語テストの分野においては、アメリカ最大のテスト作成団体である Educational Testing Service が開発した TOEFL iBT が CBT の発展に大きく寄与したことは間違いないだろう。

以上述べたテスト理論の流れを図式化したものが以下の【図１】である。



【図１】

３．CAT の意義

しかし、個々人の能力に応じて効率的に正確な測定を行うという点では、CAT は通常の CBT よりはるかに優れた特性を持つ。Brown（1997）は CAT の特徴として以下の三点を挙げている。

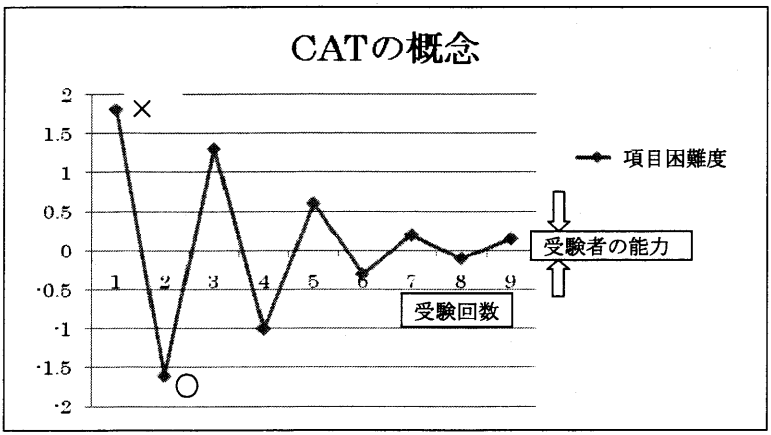
(a) the test items are selected and fitted to the individual students involved,

(b) the test is ended when the student's ability level is located, and, as a consequence,

(c) computer-adaptive tests are usually relatively short in terms of the number of items involved and the time needed

テストの項目は個々人の受験者に合ったものが選択され、テストは受験者の能力レベルが確定した時に終わり、従って通常受験する問題数が比較的少なく、時間もかからない。

つまり、IRTの理論に基づき個々の項目（問題）にそれぞれ困難度のパラメータが与えられるため、受験者の解答行動（能力）によって次に出題する問題を選択し、受験者の能力を測定するために最も相応しい難易度の問題を出題することができるわけである。従って、出題数が少なくても受験者の能力を確定した時はその問題で試験が終わるというように、効率よく正確な測定を行うことが可能だ。以上述べたCATの概念を簡略的に図式したものが以下の【図2】である。



【図2】

以下、CATで開発された科学技術英語能力試験の結果を考察しつつ、具体的な事例を通してCATの意義と問題点について論じる。

4．CATの構築

今回使用したCATはMoodleの機能を使って作成したものであり、M-CATと呼ぶことにする。M-CATは予備試験及び本試験のM-CAT実施のいずれもMoodle小テスト機能を使って行われたが、CATの部分は小テスト機能に

CAT 機能を付加した「CAT モデュール」（秋山, 2008）を使用した。構築の手順は以下の通りである。

1）アイテム（一つ一つの問題）を作成しアイテムバンクを作る。
2）これらのアイテム全ての予備試験を実施しその結果を IRT によって分析し、項目困難度の適合しないアイテムを除く。
3）適合したアイテムの中から、第1問目の問題を任意に選び、解答者に提示する。
4）2問目以降は1問前の解答行動により直近の問題と最も近い困難度の問題を出す。
5）このプロセスが繰り返され、最終的に終了条件として設定した標準誤差の数値に至った時にテストが終了する。その時の問題の困難度指数が、受験者の推定能力値となる。

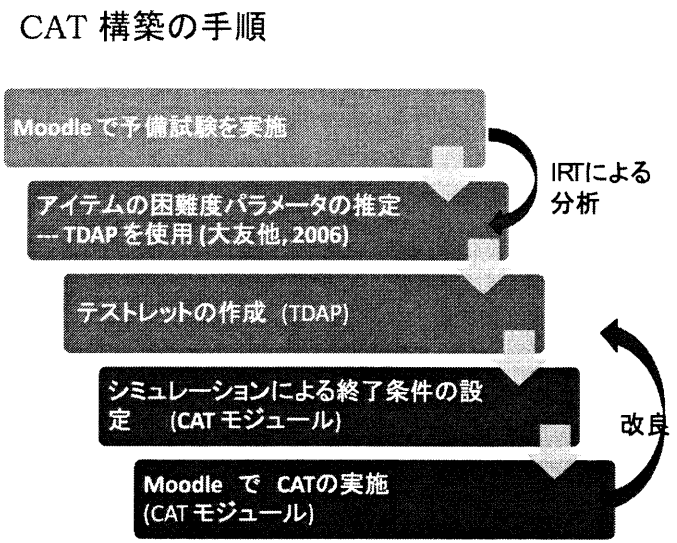
今回は、アイテムの予備試験の結果分析には IRT の中でも One-parameter Rasch Model を使った。これは受験者の数が112名と限られていることによる。またアイテムの適合性に関しては、分析の結果得られた項目困難度指数の－2から＋2の間のアイテムのみを使用し、その範囲から外れるものは除外した。さらに問題提示の方法であるが、M-CAT では一回に1問ずつの問題を出すのではなく、一番最初の回は15問、2番目からは一つのグループが3問から成るテストレットを用いた。これは1問ずつ出題するよりも項目困難度の異なるアイテムを組み合わせることによって困難度を合成することができるからである。これによりきめ細やかな困難度を実現できる。テストレットの合成困難度は TDAP（大友他, 2006）を使うことによって容易に可能であり、アイテムの組み合わせを変えて困難度を調節した上で、一つのテストレットの構成を確定することができる。また CAT モジュールを使うと終了条件も数値を設定した上でシミュレーションができる。

以下の【図3】にあるように、今回は全体で57のテストレットを作成し M-CAT を実施した。

5．実験

不適切なアイテムを除外してアイテムバンクが構築された後、上記の方法により57のテストレットが作成されたわけであるが、受験者が最初に解答す

るテストレットは15問、次回以降は一つのテストレットが３問から構成された。問題の内容は科学技術一般に関するもので、各テストレットの３問はそれぞれ、語彙、リーディング、文法から１問ずつ採用された。



【図３】

このCATの受験者は名古屋工業大学の１・２年生112名、受験時期は2009年７月、一番最初のテストレットの推定困難度は$\theta = 0$（ゼロ）即ちちょうど真ん中の困難度とした。設定した終了条件は、標準誤差が0.4である。上記４．のCAT構築の手順では最初に出題するアイテムは「任意に選ぶ」と述べた。アイテムバンクのアイテム数が多数の場合はそれが望ましいが、今回の実験では（アイテムに相当する）テストレットの数が57と極めて限られていたため、受験者の能力確定が容易になるように出発点となるテストレットは推定困難度が中位のものから始めることにした。

6．結果

まず、M-CATで受験者の推定能力値が確定するまでに受けたアイテムの数であるが、最も少ない受験者は15問、最多は42問、平均が22.7問である。TOEICの200問より遙かに少なく、学年全体で実施している統一テストの100問と比べてもかなり少ない数であることは明確である。しかし今回の実験では能力の最も高い学習者は第一番目のテストレット（15問）に全問正解することによって、能力値が＋5と最高値で確定してここで終了してしまった。これは問題の難易度と受験者の能力が適合していなかったこと、また終了条件の設定基準が不適切であったことに起因する。

第二に、受験時間はどのくらいであったかという点である。これもCATの場合は全ての学生が50分以内に、しかも90%の学生は33分以内に終了しており、平均受験時間は18.8分であった。TOEICの約2時間、統一テストの90分という試験時間よりかなり短い時間で結果が出ている。

最後にM-CATのテストとしての信頼性・妥当性を検証するために、M-CATとTOEIC、及びM-CATと統一テストの得点との相関係数をそれぞれ求めてみた。その結果、M-CATとTOEICの得点の相関係数は0.56、M-CATと統一テストの相関係数は0.65という数値が得られた。これらの相関係数から判断すると、今回作成したM-CATは一般的な英語コミュニケーション能力試験であるTOEICともやや相関が高く、一般科学技術を内容とする統一テストとはより高い相関があり、科学技術を内容とするCATとして信頼性・妥当性の高いテストであると言うことができるだろう。

以上3つの結果をまとめたものが以下の表である。

	アイテム数	時間 (min.)	CATとの相関
TOEIC	200	120	0.56
統一試験	100	90	0.65
CAT (ave.)	22.7	18.8	

7．結論

本論文では、まず第一に、CATが登場するに至ったテストの歴史的な流れを概観し、CATの今日的意義を論じた。CATはCBT一般の利点（コンピュータならではの迅速なフィードバック、マルチメディアを使った問題など）を持つ上に、さらに個々の受験者の解答行動に応じて能力に適したアイテムを出題することができる。そしてこの特徴から、結果的に時間的にも問題数においても効率的であり、かつ正確に受験者の能力を測定するという点において、古典的テスト理論に基づくテストに比べ多くの利点を持つことは明らかである。また、これらの利点は、名古屋工業大学において実施したM-CATの実験においても実証された。即ち、他の試験に比べ受験問題数は22.7問と相当少なく、また受験時間も20分未満と短かかった。かつ相関係数も一定の高さがあり、TOEICや統一試験に匹敵する信頼性・妥当性の高いテストがCATとして作成されることが示された。これらの利点は、結果的には受験者の負担の軽減、試験実施側の負担の軽減などに繋がり、CATの意義をさらに大きなものとすることは明らかである。

他方、今回の実験を通してCATのいくつかの課題が明らかになってきた。実験結果の記述にもあるように、終了条件の設定の問題と、アイテムと受験者の能力との不適合であるが、これら二つはいずれもアイテムバンクの中のアイテム数が限られていたことに由来する。今回の実験では予備テストを実施し不適切なアイテムを除いた後で残ったのは、183問であった。上記の二つの課題をクリアするためには、より多くの、しかも難易度の幅のより広いアイテムが必要である。

また今回は、一つのテストレットに語彙・文法・リーディングというスキルの異なるアイテムを入れた。テストのそもそもの妥当性を考えるとき、「何を測っているのか」という測定の対象として構成要素を抜きにしては論ずることができない。妥当性の高いCATを作成するには科学技術英語の構成要素は何であるか、という基本的な問題をもう一度検討しM-CATの目的を明らかにすることも避けては通れない。

コンピュータ・テクノロジーが発達を続ける現在CATには大きな魅力がある。しかし同時に、Dunkel（1999）が指摘するように、CATにはCAT自体が擁する幾つかの本質的な問題がある。即ち、(a) the basic principles

of assessment embodied in the CAT（CATに具体化される基本的な評価理論）、(b) the special psychometric and technical issues peculiar to the CAT as opposed to traditional or paper-and-pencil tests（伝統的な紙と鉛筆のテストに対置される、心理測定的・技術的な特殊な問題）、(c) the hardware and software used in the CAT（CATに使われるハードウェア・ソフトウェア）、(d) the administration of the CAT（CATの実施）の4点である。これら全てをクリアした意味のあるテストを作成することは決して容易ではない。これまでCATの開発がEducational Testing Serviceのような大規模なテスト開発集団によってしか行われてこなかった理由はここにある。

しかし、本論文に示されたようにMoodleのCATモジュールを使うことによってCATは現場の教員にも手の届くものになった。今後、多くの教育現場でCATを作成し実施しそして改良を重ねていく実践が積まれ、そしてそのことによって学生や生徒の能力が正確に効率的に測定され、かつCATの英語教育へのpositive wash-back effect（肯定的な波及効果）が広まっていくよう願うものである。

参考文献

秋山實. (2008). Moodleの小テスト機能をベースとしたアダプティブテストモジュールの開発, 情報教育研究集会発表要旨.

大友賢二. (1996). 項目応答理論入門, 大修館書店, 東京.

大友賢二, 中村洋一, 秋山實. (2006). TDAP 2.02 http://e-learning.ac/moodle-resources/ Retrieved September 24th, 2009.

荘島宏二郎. (2009). ニューラルテスト理論－資格試験のためのテスト標準化理論－電子情報通信学会誌, 92, 1013-1016.

Brown, J.D. (1997). Computers in language testing: present research and some future directions. *Language Learning & Technology,* Vol. 1, No. 1, 44-59.

Chalhoub-Deville, M. (2001). Language testing and technology: past and future. *Language, Learning & Technology,* Vol. 5, pp. 95-98

Dunkel, P. (Ed). (1991). *Computer assisted language learning and testing: Research issues and practice.* New York: Newbury House.

Dunkel, P. (1999). Considerations in developing or using second/foreign language proficiency computer-adaptive tests. *Language Learning & Technology Vol. 2, No. 2, pp. 77-93* http://llt.msu.edu/vol2num2/article4/ Retrieved January 10th, 2010

Edgeworth, F.Y. (1888). The Statistics of Examinations, *The Journal of the Royal Statistical Society*

Lord, F.M. (1950). Notes on comparable scales for test scores (*Research Bulletin* 50-48) „Educational Testing Service.

Spolsky, B. (1995). *Measured Words: The Development of Objective Language Teaching*, Oxford: Oxford University Press