

多様な学力を計測するためのテスト技術

池田 央*

2003年1月29日

於：東京大学大学院教育学研究科第一会議室

1 講演者紹介と趣旨説明

市川：定刻になりましたので始めたいと思います。池田先生、今日はどうもありがとうございました。これはCOEの研究会で第2回目ということになります。初めにCOEとは何かということでプロジェクトリーダーである金子先生の方からCOEについて説明をして頂きたいと思います。お願いします。

金子：今日は池田先生をお招きしての研究会です。このCOEといいますのは、文部科学省が策定しました「21世紀COEプロジェクト」というのがございまして、その中の1つとして、東京大学教育学部が所属の教官を組織し応募しまして、選定されたものであります。実は、純粋な教育学関係のCOEはここだけでございまして、心理学、生理学等を集めた人間の何とかというのが他に2、3件あります。また高等教育では1つあるということですが、純粋な教育学というCOEについてはこちらだけでして、選ばれたことが光栄であると同時に、これをきちんとやっていくことが非常に重要であると思います。

このCOEは、具体的には3つのユニットからできていて、1つは基礎学力調査分析ユニット、2番目は学校機能分析ユニット、第3番目は政策オルタナティブのユニットです。それぞれ基礎的な学力の概念などにかかわる調査、学校の機能に関わる調査、3番目は政策に関わる調査と検討を加えることになっております。現在、学力低下が新聞にも出るわけですが、それは単に今までできていたテストができなくなっている、

という単純な問題ではないのです。新しい時代にどのような能力が要求されるのか、それを単に形式的な学力テストをやるのではなくて、新しい時代に生きていく人たちが必要な能力とは何であるのか、それは必ずしも今まで測られてきた能力とは同じではないのではないかと、という問題認識が、特に私共はCOEプロジェクトで非常に重要だと思っています。私共がCOEで審査されたときに、ともするとこういうプロジェクトは、既成の学力観、既成のテストで測れるものにかたよってしまうので、それは避けるようにという注意を受けまして、それは本当にそうだろうと思います。

前々日に池田先生においで頂いたときにOECDのPIISAの方の講演をして頂きました。PIISAは国際学力調査ということで、それだけで日本では有名になりましたが、PIISAをやっている人たちは既成の学力観だけではなくてこれからの社会生活に必要なとされる能力、これからの地域社会に必要な能力とは何か、ということをもう少し幅広く考えていきたいと強調しております。それを測るためには単に既成のペーパーテストではなくて、コンピューターを使ったインタラクティブ型の計測といったものを考えるべきだと言っていました。私もそのとおりであると思います。是非、我々もそのような側面をやりたいと思っています。具体的には市川先生、南風原先生を中心にやって頂くことになると思うのですが、ヒントを頂ければということで、この分野の第一人者である池田先生にお話をうかがうことになりました。もう1つこの場で強調しておきたいことですが、COEは単に教員の組織だけではありませんので、教育学部の大学院生、学生にも活躍してもらいたいわけです。もうひとつは、大学、い

* 立教大学名誉教授

いわゆるアカデミックな研究だけではなくて、大学外でやっている会社のテストなど、様々な取り組みがすでにやられているので、そういうものと広く交流しつつ活動を進めたいと思っています。そういった点でも池田先生からサゼスチョンを頂けたらと思います。

2 講演

池田氏：池田でございます。私は大学の現役の教師はもうほとんどやっておりません。もっぱら英検のテストや、そのコンピューター化に向けての研究・開発に関わっております。そういうこともあってお声がかかったのだと思いますが、私は小中高といった初等・中等教育の問題にそれほど関わっていたわけではありまないので、コンピューターがそういった教育にいかに関立つか、関心はありますが、詳しいことはよくわかりませんので、むしろ皆さん方のほうが詳しいのではないかと考えております。

今日は、この10年間、あるいはもっと前からになりますが、わたくしがいままで考えておりますテストと、主にアメリカが中心になりますが、どんどん発達してきているテスト技術の現状を少しお話してみたいと思います。その上で日本をどうすれば良いかということについての議論はまた他の方にお任せしたいと思います。ここではいちおうパワーポイントで作りました資料を元にしながらお話させて頂きたいと思います。

多様な学力を計測するためのテスト技術

池田 央
立教大学名誉教授
教育測定研究所

1

今日のタイトルは『多様な学力を計測するためのテスト技術』ということですが、学力とは何かということをもたまたここで議論し始めますと大変なことになり、

收拾つかなくなるかもしれません。ここではむしろテストを使ってまず具体的にやれるようなこと、これをテスト技術と呼んでおきますが、そこから考えたいと思います。教育を考える際、その実体を測るものというのはほとんどテストぐらいしか考えられないと思うからです。

今日の話題

テストの役割は今日大きく変わりつつある。

- ・ 第1にテスト**目的の変化**
- ・ 第2にテスト**技術の進歩**
- ・ 第3に**CBT**の現状と課題
- ・ 第4に**教育調査**の諸問題

2

2.1 テストの役割

テスト目的の変化

- ・ 選抜のための順位付け評価（実施者のため情報）から、学習者のための**支援評価**へ（情報源の複雑化）
- ・ **多様な入学者**（出発点の不揃い）（学習者情報）
 - ガイダンス（適性、興味、レディネス、補習）、動機付け、学習ナビゲータ、コーチ、目標ガイド、成果確認と能力保証
- ・ 求められる人材の変化（OJTより**即戦力**）
 - 問われる**教育成果の保証**（何が出来るか？）
- ・ 教授学習プログラムの評価（**教育調査**）へ
 - 判断に欠かせない**マクロ情報**（評価の**標準化**）

3

最初にテストの役割についてですが、日本の場合は入学試験のイメージが頭にあって、選抜試験のことしか頭に浮かばない状況にあると思います。学校では学年末や期末試験がありますが、そういうものも先々迎える入試を想定して作られているようなところがあります。しかし、本来テストはそれ以外の目的がたくさんあるわけで、そうした目的の変化についても少し考えてみたいと思います。それからテストというのは教

育の中で最も技術化されているもので、教育的なアイデアがあればそれを具体的な形にして表すための一つの技術がテストであるといっているかも知れません。その技術についてもここ最近急速にC B T化によって、コンピューター利用の環境が整ってまいりました。しかし、その歴史的淵源はずいぶん古い話なので、その延長線上にある形として捉えていくことがC B Tの現状と課題を考える上で大事なことだと思っております。そして時間があればもう1つのテストの役割、つまり教育調査、一昨日の講演のP I S Aもその例ですが、調査の1つの道具としてもつテストの役割も大きなものがあります。それに関連して、私の知っている範囲で、アメリカで最も組織的に大掛かりなN A E Pの話をしていきたいと思います。

さて、先に申しましたように、選抜の試験というのは、順位付けのための評価というイメージが強く、それだけですと順位付けが終われば目的が終わってしまう感じがあります。今一番大事なのは、学習する人のために役に立つ評価で、そのためにはどうすればよいのか、ということだと思います。今までは、実施する人の立場から、つまり学校でいえば受け入れ側が、自分の学校に学力の高い人が欲しい、ということで入学試験を行なう。選ばれる人のことを本当に考えていたのか、ということその意識は甚だ乏しいわけで、実施者サイドが主体となったテストが中心であったことは否めないと思います。

本当に必要なのは、学習する人にとって役に立つような情報を提供できるテストということになります。しかし、それは実際考えてみると難しい。選抜や一般入試だけではなくて、一方では帰国者もいれば外国人の留学生もいる、社会人入学もあれば、推薦入学もあり、私立などでは付属高校から無試験で入ってくるような人もいます。そういう様々な人が大学に入ってくるので、昔のように全員が同一の試験を受けていて、出発点も揃い学力が一定になっていると仮定するのが難しくなっています。様々な背景をもった人たちを相手に教育をしていくためには、どうしても入ってきた人たちの出発点がどうなっているかを評価してあげないと適正な指導ができない。そういう要求が片方で生まれてきていると思います。つまり、入る前にはガイダンスをやってその人の適性とか興味とか、特にレディネスが不足している場合には、入学後も補習などが必要になってきます。

また、動機づけはどうだろうかということも大事です。特に学生の学力の低下が最近問題になってきていますが、やりたい、勉強したいという動機が乏しくなっていると言われているのです。そういう人たちを学習の始めにおいて、ある何らかの形で情報を入手する、その結果どういう学習を用意してあげるかを考えていかなければならない。それに対してサゼスチョンやガイダンスができるような情報を与えなければならぬ。いわば、学習のナビゲーターにあたるようなものですが、そういう情報が必要になってくると思います。それを短期間に調べて適切な指導をしていく、という要求が出てきている。それを具体的にテストの形で捉える必要があります。つまりテストは情報収集のための具体的な道具なのです。

学期の終わり、あるいは卒業時期になりますとこのコースを終えた人たちはどういう力がついているのか、その成果を確認できるような情報をきちんと示してやる必要があります。そういうわけで、学習の前、中頃、終わった後、テストと呼ぶかどうかは別ですが、絶えず学習を評価するような機能が授業の中に含まれていないと教育目標の達成ができないということになると思います。

一方、社会の方でも、要求される人材は昔の会社であれば、入ってくる人の素質がよければ、後の教育は会社の中でできる。オンザジョブトレーニングをしながら、会社にとって有為な人材を育成するという考え方がありました。しかし、今そのようなことは言われていません。入ってきたらすぐ明日から役に立つという人が欲しい。そういうことが要求されるようになってきたので、学校の方はそれを保証してあげるだけのエビデンス（証拠）を示してあげなければいけません。それは卒業証書というようなかたちのものではなくて、具体的にその人は何々ができるというスキルを示してあげることが、企業の側から要求されています。今日のあとの方の話題になるかもしれませんが、実際にそういったことを目指して、学校が教育プログラムを作り、コースを用意し、そして卒業生が出てきたとします。しかし、果たして意図どおりの学習者を卒業させているのか、あるいは、大学、学校がたくさんあった場合に学校全体ではどうなっているかといった情報が必要になります。経済指標のように、日本全体で経済の動きがどうなっているかを示すのと同じように、教育全体においてどうなっているかを示す指標

が必要になってきます。それが教育調査を必要とする所以だろうと思います。そういう情報を得るため、学習の初め、中頃、終わりのそれぞれにおいて必要な情報を得る1つの形としてテストが使われてくることになります。

テストに求められる情報

- ・ 教育現場でのフォーマル情報の欠如
- ・ 目標値、現在値、全体値の把握
- ・ それにはテストによる計測が必要
- ・ 全体値の把握には計測の標準化も欠かせない
- ・ 確かな位置の確認には、多種かつ精密な計測が要求される
- ・ しかも全体値情報には常に最新のものが必要
- ・ これらの要求に応えるためにコンピュータ化が生まれた。

4

テストに求められる情報が、具体的には1つの数値のような形で示されることがあり、企業などでは「数値目標」とよく言われます。企業の場合にはいろんな数値が出てきますが、学校の場合にはそれに当たるものがテストであると思います。その目標値を何にとればよいかは、現在までの状況がわかっていないと目標値を考えても意味はないわけです。その目標値を達成するには、現在はどれくらいであるか。目標値が現在値に比べてあまりにも高すぎたのでは、ないものねだりになってしまいますし、低すぎてもいけない。ちょうど体育の訓練をするときに例えば自分が100メートルを何秒で走れるかを知っていていれば、あと1秒はやく走る、2秒はやく走る、という目標が立ってきます。しかし、自分の現在値に対して、もし世界的選手のようにただ10秒台で走れという課題を出されてもできるわけがありません。そういう目標値では意味がない。テストで必要なことは、現在値を知るためというのがその1つの役割になるでしょう。

しかし、現在値の評価は全体がどうなっているかということの展望、つまり、マクロの情報もないと現在値の位置をどのように評価してよいか考えることができません。そのために、全体値の把握も必要になってきて、これが教育調査の役割になるのではないかと思います。そういう形を具体的な数値で示すのにテス

トの役割があるわけですが、そのためにはいわゆる計測、メジャーメントにあたるものが必要になってきます。それが話の前提になります。

言葉ではいろいろと言いは可能ですが、それを具体的な形で示すのが、テストによる計測ということになります。全体を比較するためにはどうしても計測の標準化が必要になってきます。後でまた触れますが、標準化がないと、出された数値の意味を理解することが困難になってくるからです。そういうふうにして標準化するわけですが、それには単純に1つのテスト、例えば知能検査の1つの値だけで全体をみるということではできません。多角的にみるとなると、これまた非常に複雑になってきますし、精密であることも要求されてきます。確かな位置の確認には、多岐でかつ精密な計測が要求されてくることになります。

それからもう1つ大事なこととして、全体値を知ることですが、全体値の基準は絶えず変わってくるものですから、一番カレントな情報が必要になってくると思います。後で触れたいと思いますが、教育情報の報告が出されるときに、すでに何年か経った後のものであれば、現状はそれよりもはるかに先をいっているわけで、それでは役に立たないわけです。つまり、リアルタイムの情報がほしいということです。そういう要求に答えるものとして、1つにコンピューターの利用が考えられます。別な言葉で言えば、IT技術の急速な進歩がそれを可能にしているということです。教育におけるコンピューター化の背景には、そうした社会からのさまざまな要求、目的に、技術的に応えるための1つの道具として発達してきたことがあるのだと思います。

2.2 採点技術の発展史

ここでテストについて振り返ると、私が学生の頃もそうでしたが、実は採点処理というのが実に変ででした。簡単に10年単位で採点技術の発達を振り返ってみましょう。客観テスト、いわゆるマルチプル・チョイス・テスト(MCT)ができたのは1910年代でした。1915年に高校の先生が考えたといわれています。選択式のもので、それにより生徒たちのチェックしたものの比較が可能になりました。どういう解答が多かったのか、少なかったのかも分かります。大勢の人の解答を比較するには画期的でした。おりしも、第一次世界

採点技術の発展史

- 1910年代 客観式テスト形式の登場
- 1920年代 偏差値(相対評価)の提案
- 1935年 IBM805自動採点機試作
- 1940年代 IBM punch card 方式
- 1950年代 アナログ式採点機械、IRT理論登場
- 1960年代 マーク式自動読取採点機(OMR)、CAI
- 1970年代 試験のコンピュータ化着手(NBME, ETS)
- 1980年代 実用化への実験的試行と改良、マイコン登場
- 1990年代 大規模コンピュータ試験の実施
- 2000年代 本格的実用化と高度化の時代

5

大戦になって何十万人という兵士たちの能力測定が必要になってきた時代で、「アーミーアルファ」とか「アーミーベータ」という形の客観式テストが登場してきました。そして、1920年代には、評価点を標準化する一つの方法として、T-scaleが提案されました。日本で偏差値と呼ばれるものの原型で、いわゆるスタンダードスコアという表示方法の始まりです。35年になると、早くもIBMが自動採点機を試作していて、これはまだ当時では実用化の段階には至りませんでしたが、アイデアとしてはそのようなものがすでにありました。

1940年代になると、ちょうど第二次大戦中でしたが、IBMのソーティングマシーンやパンチカードが出てきました。ここにおられる方はどれくらい実際にイメージされるかわかりません。若い方はほとんどご存知ないかもしれませんが、そのパンチカードに全部いちいちパンチャーが穴を開けて答案を写し取って、それをソーターにかけて集計するものもありました。私は大学生時代が1950年代でしたが、まだその頃のテストはほとんどが手採点で、知能検査の答案を採点盤という正解の位置に穴の開いた板紙を当てて、○の見える数を1つ2つと数えて、来る日も来る日もそれを勘定している作業がずっと続いておりました。IBMのマシンが終戦後初めて沖縄の米軍を通じて日本に輸入され、国勢調査の集計もそれで行けるといって話で、その偉力を目の当りにした担当官が肝をつぶしたとかいう話も聞いております。それまでは大勢のソロバンの名手が雇われて、大きな会場の中にズラッと並んで、一所懸命ソロバンをはじきながら集計していた姿も次第になくなってきたわけです。

50年代になるとアナログ型の採点機械が出てきて、これは人事院で採用されて使えるかもということをや

ったようですが、アナログですので採点の誤差が大きいため、結局は使われませんでした。しかし、アメリカのテスト界では、すでに1950年代のうちに、現在よく使われているIRT理論も登場して、フレデリック・ロードなどによる有名な論文が出てきておるわけです(その当時は潜在特性理論と言っていましたが)。

60年代になると、わたくしがちょうど大学院生の頃でした。当時、知能検査の標準化作業をしましたが、そのために約1万5千人のデータに採点盤をいちいち当てながら正答数を数えていました。私一人ではできませんから何人ものアルバイトに勘定してもらって、その採点結果をこんどは電動式計算機をまわして相関係数を出し、因子分析するという作業をやっていたわけです。これではたまらないと、わたしが計算機を勉強したいと思うようになったきっかけはこのためです。幸い、61年に渡米できて、留学先で目にしたのはETSで行なわれているマーク式の自動読取式採点機でした。今日センター試験などで広く使われておる方式です。回転するローラーの下を答案がどんどん通っていく、しかも裏表を同時に読み取って、しかもマークの濃い薄いも読み取って薄いのは消したものの、濃いのが本解答というふうに判断して採点し、磁気テープに読み取ったものをコンピューターにかけて、その集計分析をするという当時最新式のものでした。そういうことが当時すでに始まっていました。「日本もこれからはこういう時代になるな、いやならねば」という確信を持った上で、ますますコンピューターを勉強しなければならぬという思いに駆られました。それが1960年代のはじめです。

一方、当時私がいたイリノイ大学工学部でもプレート(PLATO)というCAIシステムがあって、今度は生徒がコンピューターと対話しながら勉強を進めていく、そういう研究がスタートしていました。D.L.ピッチャーという人が工学部にいて、その人たちがブラズマディスプレイ、タッチパネル式の装置を作って、小学生の生徒が出てきた問題の解答を指で押せば、それが正解かどうか分かるようなシステムができていました。60年代の中頃、私が卒業する頃には、イリノイ州の中で何百台という端末からそういう学習が実施できるようになっていました。70年代に入ると、4,000台から5,000台くらいの端末が世界に、引かれている時代になっていたと思います。

そういうふうにしてテスト解答を技術的に処理する

ことがどんどん進んできて、考え方やそれに対する理論的な裏付けも変わりました。従来の考え方では、T-scaleのように多くの能力は中レベルが多くて両端の少ない正規分布になるのだと考え、それを測るテスト結果もそうなるだろうという考え方でした。しかしC A Iの考え方では、学習すべきユニットが細かく刻まれています。細かく刻まれた各ユニットをクリアしたか、しないか、ということで次の段階に進んでいくわけです。各ユニットを100%クリアするということは事実上出来ないのですが、100%に近い形でひとつひとつのユニットをクリアしたとみなして次に進むという形を取るわけです。それを得点で考えると全部100点にならないといけないうことになる。正規分布でできる悪い下の方の人がいても仕方ないというわけではなくて、全員がそういった形で学習し、進行していくということから評価をしなければいけない。

そうした発想から「クライテリアン・レフェレンスド・メジャーメント」(CRM)という考え方が生まれてきました。従来は、集団を中心として、集団の平均を全体の基準にとってそれからどれくらい上かまたは下かといったような形で評価する。つまり平均のところが1つのノルムと考えて、そのノルムをベースとして全体の能力を表すような考え方がそれまでテストの主流でした。そういったものがCRMに対して「ノルム・レフェレンスド・メジャーメント」(NRM)と呼ばれています。それに対して、コンピューターを使いながら、ひとつひとつステップごとにそれをクリアしているか否か確かめながら次へ進んでいくというプログラム学習の考え方が、1970年代になって盛んになってまいります。

しかし、当時のコンピューターは大型計算センターによる中央処理制御システムが中心です。大きなコンピューターがあって、それから端末がタコの足のように教室に出ていて、その管理は中央制御装置で行なっていく。そのような形式では、まだ端末の数がそれほど多くあるわけではありませんから、1人1人の子どもたち全員が一斉に行なうテストはなかなかできませんでした。仮にそれを行なっても、計算機の能力がそれほど高くありませんから、集中制御的なやり方では時間もかかるし、非常に非効率です。IRTで分析をすることも当時行なわれていましたが、とても実用化の段階になってはいません。前にやった問題ができたかできないかを瞬時にして判定して、次にその人に見

合った問題を出す、という今のような適応型テストのシステムはまだ当時の計算機では無理でした。同じIRTを使うにしても、たくさんのマークシートに書かれたマークを自動読取機で読み取って、磁気テープやドラムにおとした段階でそれをIRTを使って分析し、パラメーターを決めるという形、つまり事後的な処理によるコンピューター分析が1970年代までのテストにおけるコンピューター利用と考えられていました。

ところが、アメリカではすでに1960年代にはC A Iのような「コンピューター・アシステッド・インストラクション」という形でいろんな研究が進められていました。同時に、大きな国家試験にあたるようなものについて、例えば「ナショナル・ボード・オブ・メディカル・エグザミナーズ」(NBME)という医師免許試験の記録を読むと、医師の試験委員会でも、1960年代にすでにコンピューター化の可能性を議論にのせています。「将来の医師の試験もコンピューターによって測定するようなことを考えなければいけないだろう」と。

その大きな決め手や主張のベースになったのは次のことです。医師の能力というのは知識だけではなく同時に技能、技術を伴わなければ意味がない。患者との対応力も非常に重要な要素になる。そういったものをみるためには1つの「ペイシェント・マネージメント・モデル」(PMM)が必要で、知識と同時にそれを使ってどのように判断したり、患者さんと対処していくか、ということを見るシミュレーションモデルが有益である。そういう研究をスタートさせるべきである、という話が1960年代からすでに起こっていたのです。

現在医師の国家試験はみんなコンピューター化されていますが、歴史をたどっていくと、それが完成されたのは1999年頃です。その間の30年間というのは、ひとつひとつ一歩一歩、そういう研究を実験室から始めて、実地トライアルで確かめる、患者さんがどういふふうに対応するかやってみるといったことを重ねていって、最終的に実施するようになったわけです。70年代の終わり頃から、80年代始めにかけて、1つの大きな転機が訪れたのがパーソナルコンピューターの出現でした。従来は生徒なり受験者が利用するためにはマークシートで解答しなければなりませんでしたが、しかし、パーソナルコンピューターができたことによって、それに直接解答を記入したり、パーソナルコンピューターから出てくる問題にアクションを起こして、それ

に解答するといった、個人ごとの対話式テスト形式が可能になりました。今まで、IRTのような理論が発達してきても、実際の得点計算は事後的になされていたものが、パーソナルコンピュータの能力が上がるにつれて、テストをしながら計算もやってスコアリングもできるという状況が生まれきました。

1980年代においては、そういった研究が様々なところで試みられるようになりました。もちろんETSが主要な役割を果たしておりますが、医師の試験、建築家の試験、看護師の試験、弁護士の試験、軍隊志願者用試験などさまざまな試験で実用化を目指した研究が進み、90年代に入ってそれが一挙に花開いたといえます。

90年代に入ると、続々とそのようなコンピューターテストが、しかも何十万人を対象とする、大規模試験として実施できる段階がきました。そうやって急に我々も慌ててしまったことになるのかもしれませんが、それに至るには地下で地道な研究の積み重ねがあったのであり、その結果だけをすぐ取り入れようとしても、なかなかうまくいかないだろうと思います。特に教育の場合には、教育技術をめぐる研究費が行き渡らなかったというべきか、日本の場合はほとんどテストの研究はなされていなかったと思います。2000年代に入ってから本格的な実用化、コンピューター化に向けての取り組みがなされるようになってきた感じはしますが、行く手は決して甘いものとは思っておりません。

2.3 近年のCBTの進化

CBTの進化

- ・ マーク式ペーパーテストをコンピュータ化する。
- ・ 適応型テストとIRTの応用(データベースの蓄積)
- ・ 教授学習一体型テストとマルチメディア問題
- ・ 解答構築型課題、シミュレーション・テスト(PA)
- ・ 知識工学の活用(採点と成果の査定)
- ・ 以下少し例を見ましょう。

(略)

6

そういうCBT化への道を、まずマーク式のペーパ

ーテストをコンピュータ化する、ということからみてみましょう。それは問題をコンピューター画面で提示し、選択肢もコンピューターの画面に出ていて、正しいものをマウスでクリックすればよいという形です。

これが一番簡単な形であると思います。それはアメリカの1980年代最初の頃の形です。中頃になると適応型テストも出てきました。個人ごとに問題を変えてテストするというやり方です。個人ごとに問題を変えているわけですから、隣りで受けている人の問題内容と自分が受けている内容とは独立しています。同じ領域から問題は出ていますが、内容は別の問題であるということです。しかもその解答プロセスをコンピューターは全部記録していますから、前にこの問題はできなかったの次問題はもう少し易しい問題を出しましょうとか、前の問題ができて今度の問題もできたから、次はさらにもう少し難しい問題を出しましょう、と個人ごとにアダプティブ、つまり適応させながらその人の能力に見合った問題をコンピューターが次々と出していく。そういう形式のものが出てきます。

こうした適応型テストが実用化されるには、ひとつひとつの問題に対して、これはやさしいのか、難しいのか問題ごとの特性が予め分かっている必要があるのです。そのためのデータベース化が必要になってきます。それが完成して初めて適応型テストも実用化できるようになってくる、使えるということになるわけです。

さらに進んだ形として、最近ではe-learningということが盛んに言われています。方々の企業でもe-learningの具体化に向かって進んでいるようです。e-learningの進め方は学習しながら同時にテストを受けている、テストと学習が明確に分かれていない。つまり、問題が提示されるとそれに答える、それに答えると問題が提示されるといったことの繰り返しをやっていて、コンピューターとの間に対話がなされているわけです。つまり、対話の進行によりテストを受けながら、かつ学習が進んでいく形、「カリキュラム・エンベディッド・テスト」と言いますが、カリキュラムと一体化された形のテストが登場してきます。また、コンピューターから流れてくるのも音声であったり、画像であったり、アニメーション、動画であったり、いろんなものが可能です。そのようなマルチメディアの問題を使いながら学習が進行していく形のものが出ています。これは第三世代のCBTと呼ん

でよいかもしれません。適応型が第二世代、単純択一式が第一世代ですね。

今までの選択肢型テストの1つの欠点は、文章で表現したり、問題を解決していく課題解決型の能力が、それでは測れないのではないかと、いうことでした。そういうものも、コンピューターを使って作業させることで可能になるのではないかと。つまり、解答を自分で構築していく、「コンストラクティッド・レスポンス・テスト」と呼ばれる形式の登場です。作文や技能、実技テストなどがこれに含まれます。

そういうわけで、C B T は日進月歩の形で進んでいます。日本では多くがまだ第1段階ですが、第2、3段階とつづいていかないと世界の先進国から残されてしまう危機感をもっています。とくにテストが解答構築型になったり、教授学習一体型になると、一番難しいのはその解答結果をどのように査定するか、という採点の問題であると思います。これは自分でやってみても感ずる一番の難問題で、これをどうするかについては教育工学や知識工学といった外部領域の援助を必要とする部分も多いでしょう。実際に、最近では計算機言語学やコーパス言語学などの成果を利用して、論文の査定や評価をコンピューターが自動的に行えるようになってきています。また、シミュレーションテストの結果判定にも、知識工学を利用して、コンピューターに学習機能を持たせ、人間の判断様式に次第に近づけていって自動採点を可能にするといった研究も進められています。一部のものではすでに実用化の段階に達しており、これから一層盛んになってくることが予想されます。さらに、こうした一連の動きのなかには、時間的な経過を踏まえながら、生徒たちがコンピューター上でどのように答えていくか学習のプロセスを捉える、そういったダイナミック・テストの研究も進められています。

そうした中で、C B T の具体的問題例をいくつかみていきたいと思います。一番最初は選択式の解答についてですが、今までは選択式の中にも正誤式や、摘出式、多肢選択式、組み合わせ式、順序の並替え式（再配列式）など、いろんな形式がありました。その代表的なのが多肢選択式と呼ばれるものだと思います。それは解答用紙にならんだ選択肢記号を鉛筆でマークすればよいというもので、それをそのままコンピューターに移すのがいちばん簡単です。画面上に並んだ選択肢をマウスでクリックするという形式のものです。

ここに「イギリスの首府は次のどれですか」という問題があります。シェフィールド、ロンドン、マンチェスター、エジンバラというのがあって、ロンドンをマウスでクリックすればそれで「OK」という形ですが、コンピューター化するともう少し画面を活用することができます。ここにあるイギリスの地図には、都市を白丸で示したところがあって、その中で首府はどこかと聞くことも可能です。この2つの問題は実は同じようなことを聞いているわけですが、そこで測ろうとしているものの中身はずいぶん違います。最初の方であれば地図はなしですから、位置がどこであろうと関係なく、ロンドンが首府の名前であることさえ知っていれば正解できます。しかし後の方は地図と合同させたイメージで、ロンドンがどこにあるか知っていないと正解できません、ということで明らかに測る対象が変わってきます。同じ問題を作るにしても、そこで測られるコンストラクト（構成体）が何かを考え、それと抱き合わせて実際に問題が測っているものを考えなければなりません。

多くの多肢選択問題は正解1つを選ぶようになっていますが、次の例は、水に含まれている原子が何かを問うものです。当然、水素と酸素の2つを選んだものが正解で、それ以外のものは不正解になるということです。このように正解選択肢が2つある問題もちろん出題できますが、別の問題も起こります。先ほど採点が大変である話をしましたが、ここでも採点基準の理論的な根拠をはっきりさせないと結構難しい。例えば、1つだけ正解した場合にはどう処置するのか、水素がわかったけど、酸素がわからなかったという人は半点を与えた方が能力の正しい評価につながるのか、つながらないのか、調べなくてはなりません。全部の選択肢にチェックをつけてしまった人は当然、水素と酸素の2つともチェックしていますからうっかりするとコンピューターは正解と判定するかも知れません。このように採点の基準1つを立てるにも、方針と根拠をはっきりさせないと混乱を招きます。

実際にコンピューターの中ではどのようなプログラムを書いているかという、ここに一番簡単な例を挙げます。「フランスの首府はパリである」、「この文章は、正しいか間違いか」という正誤問題（TRUE / FALSE）で、どちらかに○をつければよいという形式の問題があるとします。問題形式としては一番単純なのですが、そのことをコンピューター上のプログ

ラム、XML形式で書くとこれくらいの結構長い長さになります。多肢選択式問題の他の例を挙げるとこれよりはるかに長くなります。かなり複雑なプログラムを用意しないとできません。

これもQTI規格による問題例で、QTIについては後で説明しますが、楽器サクソフォンの図があります。選択肢をクリックすると、選択肢によって違った音が聞こえます。そのうち、サクソフォンが出す音と同じ音を出しているのはどの選択肢かというのが問題で、音を利用した問題になっています。この場合、楽器の名前は知らなくてもかまわないですね。示された楽器はどのような音を出すものなのか認知できればよいという問題になります。

また、これはスライダー形式の問題と呼びます。棒状目盛りにある針をマウスを使って動かします。すると針の数値がゼロ度から360度まで自由に動きますから、自分で正解の数値の位置で止めることが出来ます。そうすると、今までの選択肢のように三角形の内角の和が何度かということを知ること、90度とか、180度とか、360度とか、ディスクリートな数個の選択肢から答えを選ぶのではなく、自由に数値を作ることができるので、今までのようにゲスで答える危険を少なくすることが出来ます。

これは太陽系の図ですが、画面の下の方に惑星の名前がずらっと出ています。ピーナス、ジュピター、マース、...、と。これはドラッグ・アンド・ドロップ(D & D)という形式のもので、金星(ピーナス)なら、ここにマウスのポインターを当て、金星のところまで引きずって行って離せばよい、という解答の形式になります。

2.4 CBTによる出題形式の変化

以上が一番単純な問題例で、いずれも選択式の問題ですが、それを考えているうちに、もっと複雑でいろんな問題ができそうだという思いに至ります。そこで、選択式の問題から、課題式問題、記述式問題、シミュレーション問題、マルチメディア問題などへと発展できます(プログラムはそれだけ大変になりますが)。さらには、用意されたツールとか、表計算ソフトや統計ソフトを使いながら実技課題を解いてもらうとか、辞書や年表を使ってもよいとすることが出来ます。それをクリックすれば必要な単語の意味や数値が出て

きます。ツールを使うにも、今の計算機はたいいていカリキュレーターがついていますから、数値を出す必要があれば、それを使って計算すればよい。あるいは、直線や円などの図形を描くためのツールも用意されています。さらに時間的な要素を測ることも出来ます。解答時間を何分と決めてやらせることもできれば、解答時間が残り何分あるとか確認しながら、作業させることも可能になります。

そういうわけでさらにいくつかの例を挙げてみましょう。

この画面をみるといろいろな機能が用意されていることが分かります。解答をここに書くわけですが、カリキュレーターも使えるようになっているし、消しゴムを使って一度書いたものを消すことも出来ます。前の問題に戻りたい、次の問題に進みたい、ということも矢印をクリックすればできるようになっています。これは数学の問題ですが、ルーレットを回してA B C Dのうち、針がAの上にくる可能性はどれくらいか、という確率の問題になっています。

それから同じように、円周上にある3個の点を結んでできる三角形のうち、直角三角形ができるのはいくつあるか、という問題もあります。直角三角形になるのは円周角が90度になればよいのですから、斜辺が直径を通ればよいわけですね。すなわち直径を通る三角形がいくつ作れるか勘定すればよいわけですね。

このようにしていくと、自分で解答を構築していく問題をいろいろ作ることが出来ます。例えば、数式を解いたり、グラフをかかせたり、あるいは文章を要約させたり、強調すべきところとか文章の間違いを添削して自分でもっとよい文章と思うものに直すとか。自由作文であれば、フリースタイルでエッセイを書かせることも出来ますし、課題エッセイにある条件を付けて書かせることも出来ます。それからロールプレイもあって、医者免許試験の中には、模擬患者を想定したペイシエント・マネジメント・モデルなどがあります。

他にも面白いと思ったのは、現物を見たことはありませんが、「コンフリクト・レゾリューション・スキル」という問題も用意されています。これは課題解決型の問題で、2、3分ビデオが流れるわけです。そのビデオには対人葛藤場面が出てきて、途中2、3分進んだところでかなりクリティカルな場面が起こる。その時点で画面が中断され、質問が出てきて、「あなたならこ

の際どうしますか」とか、「けんかをどうやっておさめますか」という問題が出てくるわけですね。その処理の方法が4種類くらいあって、その一つを選ぶと、次の画面が出てきて、また話が少し進んでくるとまたそこで、クリティカルな画面が出てきて止まる。するとまた質問が出てくる。「こういう風になりました、あなたはこの次、どのようにしたらよいと思いますか」といった問題が出てきて続いていく。そうすると選び方によって枝分かれをしますから、話がどんどん変わっていくわけです。そのために背後に違ったシナリオをいくつか用意しておいて、それぞれの展開のたびにどう処理するか、問うことで「コンフリクト・レゾリューション・スキル」のアセスメントをするプログラムが用意できます。そういう社会心理を応用したパーソナリティテストも研究されています。

それから、「ダーマトロジカル・スキン・ディスオーダーズ・テスト」、後でその例をおみせしますが、これは皮膚科の症状の写真を見てそれを診断させるという問題例になります。その他、与えられる解答が何通りしかないというものから、付け替えで何通りもあり得る、というようなものまでいろいろな種類の問題がC B Tではできます。いくつか駆け足でみてみましょう。

例えばこれはフローチャートですが、この中に解答を入れるわけですね。画面の下にある選択肢から選んで空白に埋め込み、フローチャートを完成させる。そのフローチャートが正しければ、コンピューターが正しく動きますが、まずければそれは動かないということで、それで正しいかどうかを判定することができるわけです。

それからこれは関数 $f(x) = 2 - |x + 1|$ のグラフを書きなさい、という問題です。選択肢にある線分の組み合わせを選んで、マウスで移動し図に当てはめていくと、この解が求まるようになっています。

次にこれは積分の問題で、積分の近似値で、 $n = 1$ から3までを合計した値は、このグラフの面積のどれにあたりますか、という問題です。積分は関数グラフの下側の面積にあたりますから、そこをマウスで塗りつぶすことによって解が得られます。積分の意味がわかっているかをテストするものです。

これらはいずれも採点が割合楽なのが特徴です。

またこの図には、ある仮想国での伝染病の死者数が出ていますが、グラフの横軸の方には、年代が、縦軸にはその年の10万人あたりの死者数が書かれて

います。これにより年毎に死亡者数が減っていく様子がわかります。これをみてその理由としてどのようなことが考えられるかというのが問題です。死亡者が減少したのはなぜか、病院施設が整ってきたからか、生活環境が改善されてきたからか、その理由をめぐって考えられるさまざまな仮説を列挙してみなさいという受験者の仮説提起力を問うのが、この問題のねらいです。

これはエクセルを使った例ですが、簿記に関する具体的な数字が出ています。これに対していろんな会計計算をして、与えられた表を完成するといった実務に近い計算問題です。

この問題は、左側に季節が出ていて、右側にクリスマスとか、インディペンデンスデー、レイバーデーとかいうような祝祭日のリストが載っています。それをマウスでドラッグして、左側の季節のどこに入れるか、カレンダーの順に並べて入れ込むわけです。祝祭日の日にちを覚えているかを問う問題になっています。

それからこれは「ハイライトニング・テキスト」の問題です。文章があり、その中の特に強調すべきフレーズとかキーワード、あるいはパラグラフなど一番中心になるものを選んだりさせると、その文章の色が変わったり、チカチカと点滅したりさせることが可能です。

つぎは、文章があり、並べ替えて、一番論旨がはっきりするような文章に直すというテストです。

それに関連して最近、斎藤孝氏の「3色ボールペン」という本がありますが、それは手で線を引くという問題ですが、コンピューターなら、文章の並べ替えなども容易に出来ます。

これは、医学に関連する問題です。ここに内臓の名前がついていて、教育用にそういう臓器の学習をさせるわけですが、テストのときには名称部分は全部消えていて、ここの臓器は何であるかとか、ここは何であるかとか答えてもらうわけです。すぐに自分の答えが正解であったかどうか教えてくれ、分からなかった場合はただちに正解を教えて再学習してもらうことも可能です。

また、これは先ほど触れました皮膚疾患の問題です。コピーなので図がはっきりしませんが、お許しください。ここに示したのは実際の腕の皮膚写真とこの患者の症状を示すいろんな記述です。一方、右の方にはスクロールすると条件として、100個以上の病名が出てき

ます。このような皮膚に関する病名のリストがあって、それをチェックしていき、この写真の病気はこれにあたると指摘させます。例えば、病名は何なのか、病気の個所は写真のどこの部分か答えるのです。写真にはズーム機能がついているので、写真がはっきりしないときには自分の見やすい大きさに拡大して、どこが悪いのか指摘できるという機能もついています。

このように、シミュレーションを使ったテストの多くはビジュアルなものです。1978年頃のテストの問題は文章がほとんどでした。そのように画像というものがないのが少なかったのですが、1995年にウインドウズ95が出ると、すぐに応用され、画像で問題を解説することが容易になり、そうした問題が増えてきました。

最も進んだコンピューターテストの1つの例として、アメリカの医師の試験に「ケース・シミュレーション・テスト」があります。図でははっきり読めませんが、ここに患者さんが訪問してきたという設定があって、そのときの状況がいろいろ書いてあります。それを見ながらカルテを作っていくわけですが、その症状からどういう検査をしたらよいかとか、与えられた症状から何ではなからうかと疑い、病名を考えてそれに見合った検査をするということを行います。記入した解答によって背後にある膨大なデータベースから、それに見合った結果を出す。その出てきた結果を見ながら、今度は次のステップとしてどのような処置をしたらよいだろうか、これは再検査した方がよいか、これは処置をして入院すべきか、どういう薬を与えたらよいかなど、そういうことが進行する形になっています。最も進んだシミュレーションテストだと思います。この開発にはものすごいお金がかかっているわけです。

こちらは建築家の問題で、ホームページから例題をひいてきたものです。ここにいろんな建築設計上の条件が書かれていて、次の画面には敷地に家を建てるというモデルがあります。敷地があり、木が画かれており、そこに与えられた条件に合った家を建てなさいというような問題で、長い時間をかけて解答させるように出来ています。これはその一例ですが、ここにスーパーマーケットがあります。植木があり、パーキングロットがあり、ここにある消防署を改築したい。建物はどちら向きに建てるか、どこに入口をつけるか、事務所をどうするか...、機械設備の設置をどうするかなど。建物を決められた枠内で自由に設計させる。作図にはいろんな道具があって、消しゴムも計算器もあ

りますし、線をひいて図を書くための道具もあります。それらを使いながら解答するわけです。そしてこの図のようなものを作っていくわけですね。そして採点は、たとえば入り口が適切な場所に作られているか、出入りに差し支えないかとか、ここの位置では装置が使いにくいのではないかと、そういうことをコンピューターのほうがアクセプト、リジェクト、インディターミナント、という3段階でひとつひとつチェックして採点をするというシステムです。そういった最も進んだ建築家の統一試験などは1つのシミュレーションテストのよい例であります。

2.5 C B T 作成の困難と課題

CBT作成上の困難さ

- ・ エラーの防止 → 単純化・標準化(共通書式化)
- ・ 情報量の多さ → 圧縮化・テキスト化
- ・ フォームの不揃い → 規格化・標準化・モジュール化
- ・ 作成の生産性 → 部品化・モジュール化・テンプレート
- ・ 不正防止 → 暗号化・問題個別化・ランダム発生
- ・ 成績査定 → 解答の比較可能化・臨界点チェック

↓ ↓ ↓
かくして鍵は**集個技術**と**パフォーマンス評価**

7

そういう風にしてC B Tの作成をしていくわけですが、それを実際に開発するとなると、とても大変な手間、労力、解決しなければならない問題が山とあります。

自分で作ってみると、選択式のものはそれほど大変なこともなくできますが、そうでないものは、エラーを防止したり、また情報量が大きいですから、それをどのように整理し、圧縮するかということがあります。

フォームが作成者によって違うし、提示する場所も違ったりしますからそういったものを、やはりどうしても規格化したり、標準化したり、あるいは部品化して分けたりして、決まった形の部品を用意し、それらを組み合わせて問題の形にすることを考えないと、なかなかうまくできません。それから、テスト作成の生産性をいかにして上げるかが大事です。その他にカン

ニングの防止であるとか、違った様々な解答をどう評価していくか、いろんな困難さが出てきます。

CBT化への課題

- CATには膨大な**問題データベース**が必要。
 - 問題の自動作成化の試み
- PAになるとその**評価法**が格段に困難。
 - 知識工学を利用した自動採点システム
- 問題作成には多様な個別要求がある一方で、比較可能な**標準化規格化**が必要。
- そのためフォームの**相互運用性**を目指したQTI (Question&Test Interoperability) 企画が進行中。

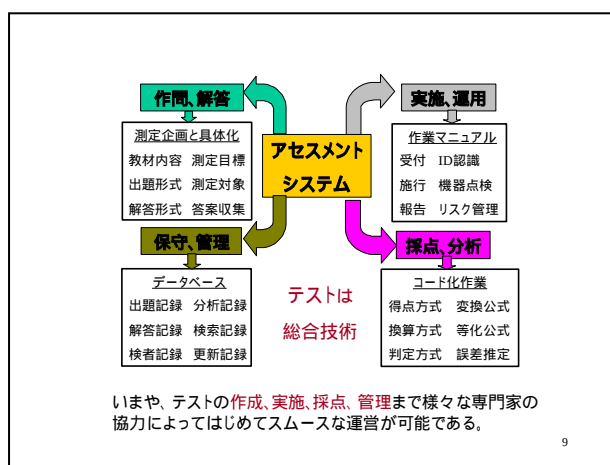
8

それらのなかで、適応型テストの問題には、どうしても大きなデータベースが必要になってきます。パラメーターを備えた問題がそうです。細かい数字はなかなか公にされませんが、医学の場合もひょっとすると十万題を超えるデータベースが用意されているかもしれません。入学試験の問題は、使い終わったらそれでおしまいです。1回ごとに使い捨てるということではなくて、長い年月をかけて蓄積されていったいわゆる標準となるような問題が、全部記録として残っていることが大事です。

それから、いまPAと呼ばれている、「パフォーマンス・アセスメント」が盛んになりつつありますが、その評価法というのも格段に難しくなります。選択式問題の場合でも、さっきのように多肢選択式で2つの選択肢が正解であると、2つのうち1つが正解であったものをどう扱うかで判断が分かれました。ましてや、PAのように多様な解答が出てきたとき、それをどう扱うかを決めるのは、全然難しさの次元が違ってくる感じがいたします。そういう研究を今後どんどん進めていかなければならないでしょう。

それから、問題作成には、こういうことがやりたい、ああいうことがやりたいということがありますから、その都度、対応を考えていては、いつまでたっても問題ができあがらないし、プログラムも書けないので、生産性が上がりません。そのため、ある程度の標準化、規格化というのは避けられないと思います。最近では問題内容では個々違っているが、フォームについては

規格化していて、いろいろな人が相互に共通利用できるようにしよう。規格に合ったフォームさえ作れば、使うコンピューターの如何にかかわらず利用できるのではないか、という考えが生まれています。それがQTIという、「クエスチョン・アンド・テスト・インターオペラビリティ」、質問とテストの相互運用可能性という概念で、現在進行中です。



9

つい一昨年に、ちょうどシアトル郊外にあるマイクロソフト社にETSのスタッフの人たちと一緒にに行ける機会があり、中身はなかなか見せてもらえないのですが、実際に、そのグループがどういうことをやっているか、会議やディスカッションにオブザーバーとして参加させてもらうことができました。そのへんの議論などを中心に少し触れてみたいと思います。これは「インストラクショナル・マネジメント・システム」(IMS)のエド・ウォーカーさんからの資料を拝借してお見せしています。例えば、テストによるアセスメント一つをとっても、たくさんの人が関わっています。1人では今のような大掛かりなことはできません。そこに携わる人々を考えると、ここが1つのリポジトリーで、こちらが問題分、こちらが受験者分のデータベースです。こちらの問題分のデータベースですが、それに関わっている人の中には、問題作成者、すなわちオーサーになって問題を作っている人がいます。それぞれの物理の問題はそれぞれの物理の専門家が作るでしょう。数学は数学です。そのような教科のオーサーです。そして、ここにサイコメトリシャンがいます。サイコメトリシャンはオーサーが作った素材を今のようなテストシステムにのせられるような形に考えてい

く人ですね。ここにあるアセッサーは、受験者の解答を判断する人で、評定者にあたります。すべての解答を完全にコンピューターで自動化して採点するのはまだ難しいですから、途中でどうしてもヒューマンジャッジメントを入れなければならないことが残ります。特にパフォーマンステストのような場合では、こういう人たちが重要な役割をします。

さて、ここに一群の受験者がおります。受験者はこの中で「解答する」という作業をしています。それを監督する人たち、試験監督者がいます。受験者から集められた解答は採点され、その記録はアドミニストレーティングのオーソリティー（試験実施責任者）によって、可否とか、免許の認定が行われます。それから、そういう作業の背後には計算機のシステム・マネジメントをする人がいます。一方、受験生の背後にはチューターがいて、試験についていろんなやり方などを教えているでしょう。そういった全部が組み合わさってテストシステムが成り立っています。今の大きなテストシステムでは、それぞれの部門で独立にやるよりも専門は違うかもしれませんが、互いに密接な連携のもと、一つのシステムとして運営を考えていかなければなりません。それを構成するのが問題を作るオーソライジングシステムであり、評価をするアセスメント・エンジンであり、ある程度のデータベースとして使えるアイテム・リポジトリです。そして、その外にそれらの運用を支えるインフラとしての巨大なアイテムバンクとか、図書館とか、教材資料が存在しています。テストシステムはその一部として重要な構成要素になっておりますが、このように教授学習全体を一つのシステムとして捉えたとき、それらをマネジメントするシステムとして、「ラーニング・マネジメント・システム」(LMS)が考えられます。そこでは各システムがうまく機能して、全体として動いていく、そういうシステムがマイクロソフトなどで考えている将来のe-learningシステムの構想であります。

私がおもしろかったのは、その中でサイコメトリシャンが重要な役割をしていることです。サイコメトリシャンはインストラクショナル・デザイナーでもあって、それはサブジェクト、つまり教科の内容だけでなく、そのよいコンテンツとテストをクリエートするだけの十分な統計学の素養のある必要がある。そういったことが記述されておることです。それから受験者のほか、監督、アセッサーつまりヒューマンジャッジ、

サーティフィケーション・オーソリティー、これは一応検定試験のようなものと考えていると思うので、試験の結果をみて資格を与えるのに値するかどうか判定する人を指しています。その人たちは判定基準も考えてそれをコンピューターに教えるという役割を持つことになるのですが、ほかにそうしたアセスメントを実際に実施するときに必要なマシンを管理するエンジニアたち、受験者たちには教科の内容を指導するチューターなどがいるのだと思います。アセスメントという仕事はそういう大勢の人たちから成り立っている大変な仕事だということです。

2.6 アメリカの教育調査

それでは後半の各種試験機関のところのテーマに話を進めたいと思います。「テスト関係研究情報」最後のページになりますがwww.imsproject.org、「インストラクショナル・マネジメント・システム・プロジェクト」(IMS)というのがございまして、そこで毎年、改訂が進んできて基準作りが行なわれております。そういうものにも目を通しておかないと世界の趨勢から取り残されるのではないかと心配しております。それを具体化していこうとしているwww.questionmark.comなどをみますと、その基準に基づいて作られたようなテスト例があり、デモンストレーションでは簡単なものしか載っておりませんが、一部様子を知ることができます。それから、ETSのリサーチレポートなどを見ると、最近の教育調査に関する研究報告などが出ていますし、それからたどっていくとTOEFLやGRE、GMATなどのそれぞれのテストでコンピューター化がどういう風になっているか見ることができます。

今まで述べたテストというのは、関心が主に個人の学力なり、ある特定の人のある教科に対する能力にあって、それを数値として評価し、集積したものを対象に考えていました。しかし、これから述べる教育調査は、アメリカの学力調査に相当するものです。目的が違って個人能力を知ることが第一目的ではなく、関心はもっぱら個々の問題、課題そのものにあり、現材の生徒たちがそういったものをどの程度どういう形で理解しているか、それをみるためのテストになっています。

教育調査 (NAEPから学ぶ)

- ・ 調査の方法 (測ろうとしているものは何?)
 - 州NAEPと全国NAEP、主調査と長期傾向調査
- ・ 解答構築型問題の重視
 - 採点者の訓練と認定
- ・ 調査の計画 (継続反復調査)
 - Matrix Sampling, BIBS Design
- ・ 比較可能な能力値の推定
 - IRT, Plausible value estimate (推算値法)の活用

10

NAEPは1969年から法律によって、議会の方から予算をもらって実施されている調査です。それは日本でいえば学力調査にあたるわけですが、特徴を最初に申し上げておきますと「ナショナル・アセスメント・オブ・エデュケーション・プログレス」で、略称をNAEPと呼んでおります。これには州単位で行なうものと、全国単位で行なうものがあります。そして「メインNAEP」と呼ばれている「主調査」と「ロングターム・トレンド・アナリシス」、「トレンドNAEP」と呼ばれているもので「長期傾向調査」を行なうものとに分かれています。ねらいはそれぞれ違うわけですが、それについては次で説明します。もう1つ重要なのはマルチプル・チョイス・テストもかなりありますが、それ以外にも先ほども挙げましたが、解答構築型の問題が多くあります。これはコンピューター化されているものも、手を使って行なうようなものもあります。作業課題を与える形のもので、理科の実験をやらせたり、あるいは音楽とか技術のように技能をみるもので、実際に歌わせたり、書かせたりする問題も含まれています。いわゆるペーパーテストで選択肢問題という形のものだけでなく、かなりのものが実際に作業したり、作文したり、実験したり、そういったことを行いながら答えるという課題が大きな特徴となっています。それには当然、採点が難しくなるわけですが、そのための採点者の訓練も充実しています。そういう教育機関を設けていて、それにパスした者に認定書を与える、という形をとっています。

それから、調査の技術ですが、これもそれ自体1つの統計的な技術として理論が発展してきました。一例が「マトリックスサンプリング」です。よく調査では層別多段抽出法とかいろんな手法がありますが、これ

はそういう受験者のサンプリングを行うと同時に、問題のサンプリングを行いながら、それをクロスさせて行き、マトリックスサンプリングという独自の方法を開発しています。それを行なう実験計画も「BIBSデザイン」、これは「バランスド・インコンプリート・ブロック・スパイラル・デザイン」というものの略で、直訳すれば「釣り合い型不完備ブロック螺旋式デザイン」というべきものです。そういうものを取り入れて、一部のデータを使いながら全体の推測ができる形の調査計画を慎重にたてております。数値にもいろんな違った種類、違った年次といったものを比較可能にするためにIRTを利用したり、「ブラウジブル・バリュー・エスティメイト・メソッド」という方法を使ったりしています。NAEPでは、すべての人が同じ問題を受けているわけではありませんから、1人1人の違った問題の解答を利用して、受けなかった問題の成績をも妥当な形でブラウジブルな値として推定する方法を研究・開発してきました。それによって、条件の違ったグループの能力を比較することを可能にしているわけです。そういった複雑な数学理論や調査技術を活用しながら全国的に行なっている調査というのがNAEPです。

表 NAEP 調査教科領域			
年	全国 NAEP		州 NAEP
	主調査	長期傾向調査	
2000	数学、理科、読解(4学年)	数学、理科、読解、作文	数学(48学年)、理科(48学年)
1999			
1998	公民、読解、作文	数学、理科、読解、作文	読解(48学年)、作文(8学年)
1997	読解(8学年)		
1996	数学、理科	数学、理科、読解、作文	数学(48学年)、理科(8学年)
1994	地理、米国史、読解	数学、理科、読解、作文	読解(48学年)
1992	数学、読解、作文	数学、理科、読解、作文	数学(48学年)、理科(48学年)
1990	数学、理科、読解	数学、理科、読解、作文	数学(8学年)
1988	公民、文書、地理、文学、数学、理科、読解	公民、数学、理科、読解、作文	
1986	コンピュータ、米国史	数学、理科、読解	
1984	文学、数学、理科、読解		
1981-82	読解、作文	読解、作文	
1979-80	数学、理科、市民/社会科	数学、理科	
1978-79	読解		
1977-78	芸術、音楽、作文		
1976-77	消費術、数学	数学	
1975-76	基本生活術、理科	理科	
1974-75	市民/社会科、数学	市民/社会科	
1973-74	芸術、基本術、読解	読解	
1972-71	生涯、読解、作文		
1971-70	数学、理科	数学、理科	
1970-69	音楽、社会科		
1969-68	文学、読解	読解	
1968-67	市民、理科、作文	理科	

National Center for Education Statistics (1999). The NAEP Guide. NCES. より

11

字が小さくて申し訳ないのですが、そのNAEPには、「主調査」と「長期傾向調査」というのがあります。

下から上の方にみていきますけれど、1969年から、いろんな評価が用意されていて、それを順繰りに4学年、8学年、12学年、すなわち9歳、13歳、17歳の生徒を対象にして、サンプリングを行ないます。2000年のデータですと、全国のサンプル学校数が2,500校、受

験生が10万6千5百人くらいです。それとは独立して、州ではそれぞれ自分の州での詳しいデータが欲しいという要望が多くなりまして、最近では州別のNAEPといったものも行なってきました。全国的な推定を行なうときには例えば、人口が多い州、それから人口の少ない州では人口に比例した形でサンプリングを行ないます。しかし、州だけの情報を得ようとすると、それでは人口の少ない州のサンプル数が少なくなってきましたから、そういう州ではもっと増やさなければいけないということになって、別枠にして州ごとにそれぞれに推定に必要なだけのサンプルを集めるという形をとっており、州では平均して225校、総数の受験者数は50万人くらいの調査を行なっております。科目として読解力、数学、理科、作文、米国史、公民、地理です。

こういう調査ですので、解釈には背景状況、バックグラウンドインフォメーションが大事です。日本ではなかなかそういったものが集めにくいと思いますが、それを集めるのに力を入れております。それが教育達成度の背景を知るのに一番重要な情報になるからです。

ここでは生徒、クラスの先生、校長先生に対して、それぞれ別の質問が行われます。生徒にはどれくらい勉強するかとか、どういった科目が好きかといったものだと思いますが、教室の先生には授業で重点を入れているのはどんなことであるとか、校長は校長で管理上注意していることは、などの質問をして、それに答えることになっております。そのような形で毎年行なっているのがNAEPです。この調査の1つの特徴は、例えば、算数なら算数の問題を出したいと思っても、1人の生徒が答えられる問題の数には限界があります。普通1人の人が受験するというと多くても100問くらいが限度でしょう。50問なら50問として全員が一斉に同じテストを受けるということであれば、個人間の比較はできるかもしれませんが、出したい問題はそれに比べればはるかに多いわけです。こういう問題も出したい、ああいう問題も出したいというのがたくさんあるわけですから、それを1人の人に全部出して解かせるというのは何日あってもできない相談ですね。そこでどのように問題群を分割して割当てるかについては、先ほどの「インコンプリート・バランスド・デザイン」を応用するわけですが、このNAEPではたくさんある問題を公平に分割して、この人にはこの問題、あの人にはあの問題、しかし、全体としてみれば

バランスがとれているような形で、問題を配分して調査する形をとっております。例えば、選択式問題が全体で200題とか300題とかあるとします。解答構築型の問題も何十題かあるとします。それらをバランスよく配分するかたちで多くの課題を与え、全体像がみえるように設計してあります。そして受けられなかった問題の成績については、ブラウジブル・バリューとして推定し、母集団全体としてどの位の成績になるかを見積もるわけです。

一方、「長期傾向調査」というのは、数学と理科、読解と作文が主で時々、公民などが入ることもあります。これらの問題は共通して非常に基礎的な問題で、ねらいは年代が変化するにしたがって、どのように基礎的な知識が動いていっているか、そのような長期のトレンドをみるための調査です。これには有名なビートンの金則というのがあって、「変化を測定するときには、測定するものさしを変化させてはならない」ということです。そのため、出ず問題は原則として同じにしておきます。共通尺度を作っておいて、いつも同じ基準で評定できるようにしておく、それが長期傾向調査のねらいです。そこで取り上げている科目も基本的な数学、理科、読解力、作文力というような基礎的なものです。これはだれもが知っておかなければならない基礎的な部分ですので、年によって必要、不必要などということが起こらないような内容の問題を集めた形式になっています。

それに対して、「主調査」の方はその時々でこのようなことを調べようというフレームワークを決めて、それに基づいて問題が作られます。だいたい、何年かおきに同じ科目が繰り返される形で進行していますが、私が注目しているのが、時代の変化に応じて重要と考えられる一回限りの特別テーマの調査です。

例えば、1986年にはコンピューターという調査科目が取り入れられております。これはコンピューターリテラシーに関する調査で、1980年代に入り、パーソナルコンピューターができて、多くの学校に入ってきました。その結果、各学校の生徒たちがどれくらいコンピューターリテラシーをマスターしているだろうか、その影響、悪影響などはどうか、そういうことを調査しようとして作られたのがこの調査になります。

また例えば、1977年から78年には「消費術」と書いているように、コンシューマースキルの能力といいえますが、意識を問うている問題が入れられています。ス

ーパーマーケットなどではいろんなものが売られていますし、テレビなどのコマーシャルがどんどん売らんがための商品宣伝をしている。そういうときに、消費者というか視聴者である子どもたちは賢くならなければならない。それに引きずられて買いたいものを何でも買ってしまうということでは困るわけで、そういったものを批判的にみる力がどれくらい備わっているかということを見る必要があるわけです。そこでこの「コンシューマースキル」の調査が行われました。

それから1976 - 77年度に行われた、「基本生活術」、これこそ、生きるための基本的な力かもしれませんが、そういう市民生活をしていくのに必要なすべて、それに対しどういう態度や考え方を子どもたちがもっているだろうか、そういうことをテーマに取り上げた調査であります。

このように、年次的に調べられている基礎科目調査のほか、その時々々の社会の要請に応じた形の調査項目も取り上げて、常にカレントな教育情報を集めているということは驚嘆に値します。しかもここで大事なことは、かなりの情報が公開されていて、われわれはそれを見ることができることです。それは、最後の方に書いてあるNCE Sの「ガバメント・ネイションズ・レポートカード」を見れば分かるのですが、いろんなサマリーレポートが入っていて、どこに何があるのか探すだけでも大変でしょう。しかし、それらを細かくみていけばいろんなことがわかつてと思います。そこで大事なことは、情報公開ということに関連してくる問題だと思っています。

アメリカの多くの教育データやテスト結果が、いろいろなホームページに出ております。たとえばTOEFLなどにしても、1993年から2002年までのデータの平均点をずっとみてみると、この青いのが全世界の受験者平均点ですが、少しずつ上昇していることが分かります。これは「スケールドリフト」と呼ばれる現象でしょう。同じ共通尺度で等化して行くと、段々甘くなっていくので、わずかずつ上昇が見られます。これに対して、各国の平均点はどうでしょう。日本が一番最下位のクラスで、それは前も今も変わりません。相対的にはやや下がっているといえるでしょうか。黄色は中国のグラフですが、初めの頃は世界平均に近かったのですが、だんだん力を伸ばし今では世界平均を超えています。韓国はどうかというと、93年頃は中国よりも日本の方に近かったのですが、英語教育を重視す

るようになってその差は開き、今では中国に近づこうとしています。

TOEFL自身は留学試験のデータですが、こういうのは教育政策を考える上で、全国的な指標としてのものにも役立つことがよくわかります。

なぜここにこれを出したかということ、情報公開することによって、テストデータが我々に見える形になっていることがわかると同時に、もう一つ大事なことで、IRTなどのテスト理論を使って共通化された尺度を用いることにより、こうした変化が比較可能な目に見える形で示されているということです。

例えば、SATにしてもそうです。この図はアメリカの大学入試で使われるSATのテストで、少し古いデータですが、1950年代から80年代までの図を表しています。よく知られているように、1950年代から60年代にかけて、平均点はほぼ横ばいに平行しておりますが、1960年代から80年代にかけては確実に低下して、平均点が下がっています。これは有名な「ネイション・アット・リスク」などのレポートにも引用されるわけですが、こういうふうな60年代後半のベトナム戦争後から70年代にかけてアメリカの学力が低下しているのではないかと教育危機が叫ばれました。その原因についてはいろいろ議論があるでしょうが、とにかく事実としてこのような低下現象がみられるということは、重大問題であるということになって、教育改革が進められたわけです。そういう貴重なデータなのです。これも共通尺度を使用しているからこそ分かることでした。

2.7 わが国の問題と展望

わが国での問題と要望

全米教育調査を見て思うこと

・ 情報公開と資料の共同利用

教育調査は多大な税金を投入して継続的に行うものであり、そこから得られた情報は国民すべての共有財産として還元しなければならない、そのための法制度も必要である。

・ 高度調査測定技術の導入

教育調査は学力の持つきわめて複雑な要素を解決する必要がある、それには調査結果を多角的に捉えることのできる高度な測定技術が必要である、そのための専門家養成も急務である。

・ 即時処理を可能にする集個システムの開発

集計調査は個別情報を集めるための集個技術が完成してはじめて可能である。そのため各学校に配備されつつある通信ネットワークの活用を考えなければならない。そうできれば1年間のサイクルの間にし得るアウトプットは貧弱きわまるものとなる。

NAEPの「コンピュータ自動報告書作成システム」に類するようなものが必要である。

私にとって重要なのは、このように比較可能な形に示された事実が調査に生かされているということです。「センター試験」のホームページに出ているものをコピーして、主要科目の平均点を平成9年から並べてみましたけれども、これをみただけでは変動の中の長期傾向というのは読み取れないというか、ランダム変動としか言いようがなく、よくわからないですね。ここで言いたいのは、教育調査資料は可能な限り情報公開することによって共同利用ができるようにすること。つまり何か大きな学力調査をして、多くのデータが集まったとしても、共同利用できないのではそのデータが活用されず死んでしまうこと。特に継続調査のような場合、実施する一機関がいくら頑張ったとしても、その中の限られたスタッフと予算では、折角集められた情報の僅かしか利用できず、貴重な資料が社会みんなの共有財産にならないことです。そこでは限られたことしかできない。例えば、極端な場合、何十万人という調査をやって、その各問題の正答率を出しただけで1年が過ぎてしまって、次の調査がまわって来てしまう。その繰り返しで、それ以上の深い分析が、いつまで経っても手付かずにならなくなってしまふ。きちんと枠組みを決めて、そのデータを広く教育研究者に渡して分析を任せれば、いろんな角度からのさまざまな分析が進み、教育研究もずいぶん進歩するだろうと思うわけです。もし研究者が個別に僅かな研究費をもらっても、そこで集められる情報はローカルなデータでしかなく、これまた一生かけても一般化できるだけの十分な資料は得られないということです。こういう教育指標として役立つような資料は国家的な規模のもとでしか得られないので、できればそういうものを用意してもらいたい。そして、それが得られたらそのデータをいろんな研究者がいろんな角度から分析できるように公開してほしい。そういう気がするわけです。

それともう1つ重要なことは、NAEPにしてもTOEFLやSATの報告にしても、いわゆるサイコメトリックスとか、教育調査法の高度な技術が使われているわけです。共通尺度で比較することが可能になるようなシステムを構築することは不可能なのではなくて、理論としてもあるわけですから、そういった道具の使えるインフラの構築が必要なのです。そのためには膨大なデータを短時間でシステムチックに集められる、そういう技術も必要です。私はそれを集個システム、あるいは集個技術と呼んでいるのですが、このよ

うな研究を進めていくにはどうしてもそういったシステムの構築が必要になってくるだろうと思います。

結果の生かし方から学ぶこと

情報公開とNAEPの報告書

- NAEP Report Card(政策立案者向け)
(<http://nces.ed.gov/nationsreportcard>)
- Highlights Reports (親、教師、教育関係者等からのFAQ)
- Instructional Reports (教師、教材開発者に役立つ資料)
- State Reports (参加州の学校、教育行政官向け統計資料)
- Cross-State Data Compendia (州間比較のための要約統計)
- Trends Reports (数・理・読・作の長期変化と傾向パタンの要約)
- Focused Reports (より広い教育的意味を探索したレポート)
- Summary Data Table (背景情報と成績との関連分析情報)
- Technical Reports (作成から分析までの専門家向け技術報告)

13

今日の話と直接の関係はないのですが、お配りした一つのプリントと関連して、集個システムの重要性について最後にお話ししたいと思います。今すぐの可能性はないにしても、どうしても政策的に考えて頂きたい、というつもりで書いたものです。まず情報公開とNAEPに関連することですが、情報公開で一番問題になり、みんなの議論になるのは、個人データを扱う以上、プライバシーの問題やセキュリティの問題がからんでくることだと思います。NAEPで行なっている方法で気づくことは、読む人の対象を考えて、いろんなタイプの報告書が出ていることです。先ほどのホームページでみられるNAEPの「リポートカード」というのは、主として政策立案者に役に立つようなデータではないかと思われます。そのほか、有料かもしれませんが、中には資格が問われるものもあるかもしれませんが、申し込めば手に入るのではないかとと思われる報告書がたくさんあります。親や教師や教育関係者からのいろんな質問に対する解答が出ている「ハイライトリポート」、教師や教材を開発している人たちの参考になるような資料の出ている「インストラクショナルリポート」、それから参加した州について学校とか教育行政者向けの統計資料としての「ステートリポート」、異なった州の比較ができるような比較可能な「クロスステート・データ・コンペディア」、といった要約統計が出ているようです。

それから、年代別の傾向がわかるような「トレンドリポート」、数学、理科、読み書き作文などの長期変化

をみていくレポートです。「フォーカスドリポート」はより広い教育的な意味をデータから見出すような、そういう議論に役立てるレポートですね。それから背景情報との関連をみたいときの「サマリーレポート」とか、研究者に役立つような問題作成から分析までのテクニカルな話をまとめている「テクニカルレポート」とかいろんな形のレポートが出ています。そういうものは大きな組織でないとなかなかできないことですが、その重要性を認め、莫大な実行分析予算をつけて大々的に行なっているのがアメリカのNAEPです。今のところ一部でしかありませんが、例えば、PISSAのような調査、あるいは最近行われたTIMSSなどの調査とリンクさせてそれと比較できるようなデータも視野に入れていると聞きました。

わが国の問題点

- **教育の基礎インフラ**
 - 教材、参考資料、項目データベース不足
- **情報技術化の遅れ**
 - 即時集個技術、情報検索、表示技術
- **教育研究開発技術者の育成**
 - Instructional designer(ID),
Scenario (Item) Writer, Test Analyst

14

それでわが国の問題点ということになりますが、こういった諸外国の状況を考えてみたときにいくつか感じることは、1つは先ほどのようなデータベースとか教材、それに対する参考資料、そういったものが非常に不足していることです。基礎インフラが不十分である。前に新聞で見ましたが、総合科目で自主的に勉強させようと思ったら、学校の図書室に不足している図書が6,000万冊あるという記事があったように思います。そういったものを増刷というのは大変なので文部科学省が何かで、1つずつファイル化して、データベースにし、全国の学校からダウンロードできるように、共同利用組織を作ってはどうかと思うのです。そうした発想は無理でしょうか。

また、いろいろなテスト問題について、この問題の難しさがどのくらいだ、この問題は全国で何%くらい

の人ができるのか、といった情報はまったくありません。書店に学習参考書はたくさんありますが、この問題ができる人はいったい全国でどのくらいいるのだろう、と私はいつも気になるのですが、どの参考書にもそうした情報が集められて資料が載っているという話は聞いたことがありません。項目データベースが欲しいという気がするのはそのためです。そのためには、そういうインタラクティブな機能をもったネットワークシステムのようなものがあって、著者と読者、問題提供者と解答者の間で、常時そのような情報が集められたりフィードバックされたりしないとできないでしょう。

銀行のATMとか、セブンイレブンのPOSなどは経済の中での情報集個技術が発達した例であると思います。セブンイレブンであれば全国で1万店くらいの店で、どういう商品がどれだけ売れているかということがリアルタイムで分かります。個々の店で何が欠けていて、すぐ補充しなければならないのは何か、常時オンラインでわかっているという状況です。銀行でもATMでリアルタイムでお金の出し入れができる。ま、時にいろいろトラブルも起こりますが、そういう集個システムのおかげで、航空機の座席予約であったり、ホテルの予約であったり、求めればすぐどのホテルが空いているか、どこの列車が空いていてすぐ乗れるかなどの情報が得られるわけです。そういう面でいえば、即時にということでもなくとも、少なくとも一年単位くらいで教育の現状がわかるようなシステムがあれば、日本の小学生はこのくらいのレベルの問題や課題なら今このくらいまでできるようになっているとか、そういったことが把握できるものがあるというのはどうでしょうか。それが二番目の問題です。

最後に、もう1つ感じることは教授研究開発技術者というのが日本で非常に不足している。インストラクショナル・デザイナー、IDといわれる人たちです。私が今、心配しているのはe-learningです。これが急速に盛んになってきていて、多くの企業がe-learning化に向けて突っ走っています。しかし、エンジニア的な発想でものを考えて、教育学的な配慮など欠ける場合がある。e-learningの教材を作れば、学習のたびに、すぐテストしなければいけないことも起こる。その場合でも、問題を出してそれに○か×かを答えればそれで済む、という話で終わってはいけません。そういうことからすると、片方でどんどん技術が進んでいくとい

うのに、中身の検討の方は取り残されて、ある重要なものが欠落したまま進んでいってしまう。そういう危険を強く感じております。

例えば、先ほどのお医者さんのシミュレーションの例ですが、いま、ある症状を持った一人の仮想的患者さんが来院して、指定した検査を受けて、結果を見て、その結果からどのような病気が診断されるか、それに対しての処置はどうか、処置の結果は正しかったのか、そういうことを判定する過程は1つのストーリーをもったシナリオ制作であるわけです。シナリオまで考えてプログラムを立てないといけない。ということは、今までの単独のアイテムだけ、問題だけを考えていけばよいというテスト作りよりもはるかに難しい問題を抱えているわけで、そういうことを考える専門家というのが必要になってきます。それぞれの教科で必要になってくるし、それから得られたテスト結果をどう読み取るかという、テスト・アナリストと呼ばれるような職種の人もこれから必要になってくるということです。向こうで聞いた話ですが、いま専門のサイコメトリシャンがアメリカで千人か2千人くらいだそうですけど、2015年までには、ひと桁上がって1万5千人か2万人くらいは必要だろうという話をしている人がいたように思います。そういうことを考えると、インフラの問題や情報技術の普及と同時に、それを育て、開発していく若い人が増えて行って欲しいという気がいたします。

Web Siteから世界を見れば

検索用語 (yahoo.com より)	検索数
Glossary of Psychological Measurement	63,900
Glossary of Educational Assessment	139,000
Glossary of Test Terms	104,000
Glossary of Statistical Terms	159,000
Computer Based Counseling	492,000
Computer Assisted Career Guidance	149,000
Computerized Adaptive Testing	20,800
Multiple Choice Type Test	1,850,000
Constructed Response Items	518,000
Simulation Test Items	454,000
Automatic Item Generation, test	243,000
Automatic Scoring Systems	159,000
Latent Semantic Analysis	26,400
Automatic Essay Scoring	5,170
ISO, Educational Assessment	50,600
Question & Test Interoperability	68,800

15

ついでですが、私がyahoo.comで、「Glossary of Psychological Measurement」などという言葉でウェブサイトをひきますといろんなことが出てまいります。

例えば、「Psychological Measurement」で6万3千9百件、「Glossary of Educational Assessment」が13万9千件、「Computer-Based Counseling」が49万件です。日本でカウンセリングというと1対1で、クライアントとカウンセラーの対話みたいなものがイメージされますが、実りある対話を支える背景としてコンピューター支援による情報提供の役割がきわめて重要になってくると思います。こういう49万2千件も出てくるという数の大きさは、そうした事態の反映ではないかと思うのです。「Computer Assisted Career Guidance」で引くと、14万9千件が出てきます。なお、これは当然かもしれませんが、「Multiple Choice Type Test」でみてみますと、185万件が出てまいりました。

最近おもしろいのは問題を自動的につくってしまうという試みです。先ほどアイテムプールが重要である、決め手であると申しましたが、実際にアメリカで聞いても、項目をどんどん作ってアイテムプールに加えていくというのでは間に合わない。だから、この先は問題を自動的につくって行くシステムが必要であるということでした。ETSの先端グループがやっておりますが、例えば「Automatic Item Generation」、そこでは、数学の問題などはフレーム、問題の枠組みですね。どういうことをさせたい、問いたい枠組みを指定すれば、問題で使われる数値などはコンピューターが自動的に決めていくといった問題作りのシステム開発が進められております。

スコアリングの方も「Automatic Scoring Systems」で引くと15万9千件も出てきました。論文を含めての自動採点システムですが、いろんな作業課題を自動的に採点していくという採点システムの研究開発で、いくつかのグループで進められ、すでに実用化されているものもあります。コロラド大学などで開発された「Latent Semantic Analysis」は、エッセイの意味解析を行なって、評定を自動化しようとするものですし、ETSで開発されたe-raterのように、すでに大規模試験で実用化されているものもあります。

CBTの標準化に対しては、ISOという「国際ナショナル・オーガニゼーション・フォー・スタンダード化」という組織があります。いわゆる工業規格の相互認定システムで、日本も加盟しているところも多いと思います。製品に対して標準規格の条件を満足しているかどうかということを認証する機関です。

そういったところでもテストをその1つとして取り上げようという動きに向かっております。ですから日本国内で使っていてもそれが世界的に通用するかというと、ISOが認める基準を満たさないといけないということにやがてなるかもしれません。例えば、アイテムのフォーマットや、採点システムが世界的に認証されるシステムの条件を満たしていないテストは国際社会の中では通用せず、相手にされない時代がくるかも知れません。私はそんな予感がしているのです。日本には日本の独自のいき方があるかもしれませんが、それと同時に、今の国際化の中で外国を相手に、外国からの留学生を受け入れて、それを教育するという際に、今の形のままで通用するかどうかはいずれ問題になってくるのではないかという気がしています。ここで十分に示すことはできませんし、ホームページに出ているものに限られますが、いくつか書いておきましたので、それをたどっていけば、いろんな情報が出てくると思います。

各種試験機関ホームページ

- <http://www.toefl.org> TOEFL
- <http://www.ets.org> ETS
- <http://www.act.org> ACT
- <http://www.ericac.net> Eric
- <http://www.are-solutions.com> 建築士試験
- <http://www.chaucey.com> 資格試験各種
- <http://www.collegeboard.com/ap/students/apcd/html/apintro.html> collegeboard試験デモ
- <http://www.knowledge-technologies.com> 論文試験
- <http://www.intellimetric.com> 論文試験

16

C B T の実際の問題例をみるのは限られていますが、私が気づいたのでは、collegeboardで出しているAPという「アドバンスド・プレズメント」にはいろんなデモ用問題があって、受験生はこの勉強をしてくださいという表があります。数学などは、マウスを使ってできる逆関数のおもしろいデモなどがありますし、お金を出して、関連する参考書とか、CD-ROMなど買えばいろいろと実際の問題例が出てくるかもしれません。

また、intellimetric.comをみると例えば、ペンシルバニア州は年間140万人が受験して自分の作文を評価

してもらっているというような記事も出ていました。

日本でのCBTテスト例

- <http://www.alc.co.jp/eng/index.html> アルク
- <http://www.evidus.com> CAT英語テスト
- <http://www.globalenglish.com> 英語学習
- <http://www.englishlistening.com> 聞取学習
- <http://www.ordinate.com> phonepass 電話

17

さて日本のC B Tも、最近は少し出てくるようになりました。私は日本の現状をあまり知らないのですが、アルクなどでも、コンピューター上で受けられるテストがあります。わたしが関係しているのではcasec.evidus.comでひけば適応型英語テストで、リスニング、ボキャブラリー、ユーセージ、ディクテーションの4領域の技能を測るものがあります。今までに、10万人くらいの受験者が受けていると思います。

globalenglishとかenglishlisteningをひくと、本部はアメリカですが、日本語の文字や説明も出ていて、英語のリスニングができるようになっています。

ご存知かもしれませんが、最近ではOrdnateという会社が作ったPhonePassという電話で発音を診断してくれるものもあり、日本ではトムソンラーニングが来て盛んにやっています。携帯電話からできます。もちろんお金はとられますが、申し込みして受験票をもらって、受験番号をもらったら、それをもとに電話すると相手がいろんな問題を出してきます。一番簡単な問題は向こうが「これから英語の単語を言いますからそれと同じものを言い返してください」というものです。

それを真似しながら答えると、音声認識装置を使って、こちらの発音の波形と向こう側の波形とを比べて0から8くらいまでの段階で採点してくれます。10分間とか5分間のテストがあります。

日本のテストのC B T化は、これからだと思いますが、アメリカの資格試験などと比べると、先はまだ長いという気が致します。そういうテスト関係の研究機関のアドレスはここにありますが、ひいてみるといろいろな情報が出てきます。U C L Aのグループは

テスト関係研究情報

<http://www.cse.ucla.edu>

<http://www.ncme.org>

<http://www.imsproject.org>

<http://www.questionmark.com>

<http://www.ets.org/research>

<http://www.eschoolnews.org>

<http://nces.ed.gov/nationsreportcard>

18

ベーカーさんがリーダーになってテストスタンダードを作っています。そこではいろんなパフォーマンスアセスメントの評価に関する研究を進めていて、それに関する多くの関連論文を見ることができます。それから、「ナショナル・カウンシル・オン・メジャーメント・イン・エデュケーション」、略称NCMEというのがあり、ここはテストに関連する研究で、最近どういうテーマのものがあるか知ることが出来ます。QTIに関しては先にあげました。eschoolnewsはe-learningをやっている学校の紹介、新聞、パソコン上で新聞記事や調査結果を読むことができます。専門雑誌の論文ですと、どんなに早くても1年遅れ、普通なら2年遅れでしか研究情報が得られませんが、最近ではインターネットなどを使った情報がどんどん増えています。今から40年以上前、私が最初にマークシートによるコンピューター採点をみたときのカルチャーショックに続けて、それ以上のカルチャーショックを受けている感じです。

取り留めのないお話を長々としてしまいましたが、時間もオーバーしましたのでこれで終りに致したいと思います。

3 質疑応答および謝辞

市川：ありがとうございます。池田先生から最近のテスト事情をこれだけ詳しくうかがえたので、私たちはこれから学力調査を考えるときにぜひ生かしたいと思っています。今日は残念ながらあまり質問の時間がとれないのですが、またこれから時々、池田先生にはさらにうかがうことがあると思います。いろんなこ

とが動いているということがよくわかりました。いろいろ質問などあると思いますが、また池田先生とは長いお付き合いをさせて頂くということで、よろしくお願い致します。

南風原：池田先生は、民間企業などかなり指導されていると思うので伺います。例えば、マルチメディアを使ったテストを作る、それはものすごく大変だということでしたが、それを請け負って中心にやるといったことは…。

池田氏：それは若い人たちが自分のテーマとしてやって頂けるとありがたいですね。

南風原：企業としてそういったものを請け負っているところはあるのですか？

池田氏：企業もそれだけの人が足りないのです。やろうと思えば、いくらでも手を広げることができるのですが、いま引き受けてもね。

市川：実際にそういう予算もあって動かそうとしている会社は、日本でもたくさんあるのですか？

池田氏：NTTなどもやろうとしています。そこはネットワークのほう为中心で、それにのせられるテストが欲しいわけです。向こうもどこかがコンテンツを提供してくれることを欲しがっています。

市川：つまり、人がいればやりたいということですね。

池田氏：そうです。それとどうやっていくのかというのがわからないということです。企業でも、国内では特に外資系がすごいですね。企業内教育のために。薬を新しく開発しても、昔と違って世界中に人が散らばっていますから、それを集めて新薬の効能とか使い方を研修させることがもうできない。それで、パッと教材を作って全部ばらまいてそこで勉強してもらう。日本の場合はいままでの蓄積がないですから、急に慌てて他のところがやっているから、ということで急にやっても泥縄式のところがあるのです。それで企業だけが突っ走ると心配で、これでは教材を仮に作ったとしても、「こんなものではつまらないではないか」「こんなもの中途半端だよ」と飽きられてしまうとそれでおしまい、という危険性はあると思います。

市川：そういう点では、若い方もがんばって頂きたいと思います。ありがとうございました。