

第12回研究会

全米学力調査 (NAEP) 概説

—テストデザインと統計手法について—

村木英治 (東北大学教授)

2005年11月24日

於：東京大学赤門総合研究棟 A200講義室

NAEP (全米学力調査; the National Assessment of Educational Progress) は、唯一、全米規模で定期的に行われている教育測定である。その時々に関心の高い教育テーマについての到達レベルを計測するねらいの他に、1960年代からの長期的なトレンド推移、各州別の到達度の状況などのレポートが出され、多様なレベルでアメリカの教育政策に大きな影響を与えている。また、実施機関である ETS (Educational Testing Service) では、NAEP の目的を適切かつ効率的に実施するために、様々な教育測定の技術革新を生み出している。講演者は、ETS で 11 年間、心理測定分野のエキスパートとして NAEP に携わってきた。本講演では、NAEP の概要紹介の他、教育測定技術の革新について IRT (Item Response Theory, 項目反応理論) との関わりを中心に解説した。

(本稿は講演内容および事後の質疑応答内容を事務局が要約的にまとめたものである。)

1 はじめに

村木: 本日は、お招きいただきありがとうございます。私がアメリカから日本に帰国したのは4年前です。帰国前の1年間は ACT (American College Testing) におりましたが、それ以前の11年間は ETS で NAEP の仕事をしてきました。ちょうど私がシカゴ大の Lord の下で博士課程を修了した頃に NAEP のプロジェクトが ETS に移りました。時を同じくして、私も ETS に行く機会を得ました。ですから、NAEP と共に動いていると言えるかもしれません。

2 NAEP とは

NAEP とは、ご存じのようにアメリカで定期的に行われる唯一の全国規模の教育測定で、アメリカ全土の就学年令の児童生徒を対象にしています。読解力、作文力、数学、科学、歴史などの各分野の学力を測ります。現在

は、そのほとんどを ETS が行っておりますが、初期の段階では全米から多数の著名なサイコメトリシャン(心理測定専門家)達が動員されました。

アメリカではご存じのようにそれぞれの州でアセスメントが行われていますが、そのアセスメントの大元になるのが NAEP です。ですから、各州で行われている独自のアセスメントも NAEP に連動しながら動いています。各州においてもサイコメトリシャンが働いていますから、NAEP で仕事をするということは全米のサイコメトリックに対して影響力を持つわけです。

ETS の中で NAEP に関係している専門家には3種類あります。1つは心理測定・統計関係のグループ。それから純然たるデータ・アナリシスを担当するコンピュータ・プログラムの専門家グループ。さらに、実施の運営管理をするグループです。大体、1つのチームにこの3種の専門家が最低1人ずつ入り、そこにヘルピング・スタッフが加わる形です。ETS の中に NAEP に関与する人たちというのが大所帯で存在しています。

■心理測定学における新しい技術を導入

NAEP が優れているのは、心理測定学の新しい成果をどんどん取り入れてきた点です。ETS に NAEP のプロジェクトが移ったときに IRT を使い始めたのですが、それ

を機に Robert Mislevy の Plausible Value や、Popham の Item Sampling など、心理測定学における新しい技術が旺盛に取り入れられました。ですから「測定した「結果」についてはもちろん、「手法」についてなど、技術的な観点からも注目されている調査です。

■ 3つのNAEP

NAEP には、大きく 3 種類あります。1 つは Main NAEP です。これは、その時々で教育的に重要な関心事を重点的に測る調査です。

もう 1 つは、Long-term Trend NAEP です。長年にわたり同じテスト項目を使って調査を行い、学力が下がっているのか上がっているのかを確かめることをねらいとするものです。Main NAEP においても年度間のリンクは貼っていますが、より信頼性のある Long-term Trend は、毎回、同じ項目を使って行われているこの調査で測られます。

それから State NAEP。全ての州が行っているわけではなく、参加していない州もあります。Main NAEP にリンクさせながら各州の学力状況をしっかりと把握したい、というねらいのものです。

■ 垂直等化(Vertical Equating)

また、あまり知られていませんが 4 学年、8 学年、12

学年（9 歳、13 歳、17 歳）の間でテスト結果の垂直等化（vertical equating）がなされています。この 3 つの学年の問題の中にリンクのための共通項目(common item)が入っています。同じ被験者を 4 年おきにフォローして調査する方法ではありません。

しかしながら、公開されている NAEP のレポートを見ていただければお分かりのように、この点について強調されることはほとんどありません。理由としては、垂直等化を強調するには共通項目がやや少なすぎるということもありますし、そもそも垂直等化をする必要性がどの程度あるかも明確ではありません。年代別の発達を検証するのは NAEP の目的として強くは意識されていないようです。従って、垂直等化の件も、あまり強調されていません。

■ 調査対象

全米のおよそ 2,000 の学校から、約 10 万人の学生がサンプリングされます。悉皆調査ではありません。

アメリカが日本と異なる点は、様々なエスニシティ・グループが存在していることです。白人や黒人、アメリカン・インディアンや東洋人など。また、東洋人にも色々あります。そして、それによって居住地域も違います。こうしたエスニシティによる違いを調べることが NAEP の非常に重要な目的になっています。

したがって、調査のターゲットとなるグループを明確にし、そこから十分な標本を得るような計画をします。そのようなねらいに即してかなり限定的な設計をしている部分があります。

■ 心理測定の専門家として

先ほど、ETS の NAEP のグループには心理測定の専門家が入るとお伝えしましたが、私も専門は心理測定で、

スライド 1

National Assessment of Educational Progress (NAEP)

- NAEPは定期的に行われている全国規模の教育測定としては唯一のもので、アメリカ全土の就学年令の児童生徒を対象に、読解力、作文力、数学、科学、歴史、その他の分野での学力を客観的に測定するために、1969年から始まった。
- 州独自のアセスメントからのリンク
- 心理測定学における新しい技術の導入
- Main NAEP 主調査
 - 調査時点における全国的な学力傾向の調査
- Long-term Trend NAEP 長期傾向調査
 - 長期的な学力変化の調査
- State NAEP 州調査



Polytomous IRT (多値項目反応理論) が専門でした。最初に担当したのはライティングの評価です。その後、米国史やリーディングの担当も行いました。

私を含めサイコメトリシャン (心理測定専門家) 達は、NAEP の担当を通じて、全国調査の結果に責任を持つことはもちろんのこと、一般性のある理論や方法論を開発するという意気込みも同時に持って取り組んでいました。

といいますのも、先述のとおり、ETS に NAEP のプロジェクトが移った際に、IRT の導入も行ったのです。IRT は単純なトータルスコアによる評価などに比べると、少しややこしい。時折、項目によっては、パラメータが推定できない、または教科によって違いがある、あるいはブロック内で後ろに配置されている項目の反応が出にくい、など一筋縄ではいかない問題が色々起こるわけです。それゆえ、博士号を持っている専門家が各教科に対して研究作業も兼ねながら責任を持って行っているわけです。

私たち、サイコメトリシャンは、それらの課題に対応する方法論を開発すると同時に、自分の担当する教科の分析を行います。後ほどご説明します項目(item)の分析、calibration、Plausible Value の算出など、一連の全ての流れに責任を負うのが、ETS のサイコメトリシャンの仕事なのです。

また、ここにご紹介いたします”NAEP Technical Report” 非常に分厚い本です。NAEP で用いられた技術革新やテクニカル・メソッドの大部分がここに書かれています。これを、2年に一遍ぐらい発行するのもやはり私達の仕事です。

そして、この度、この本をベースに、東北大の荒井先生が中心となって「全米学力調査 (NAEP) の研究」¹⁾ と言う報告書を作りました。ねらいは、NAEP の学力テストで、どのようなメソッドロジーが使われているのかを日本に紹介しようということです。

■個人の学力を報告する必要はない

NAEP は、サンプリング調査ですから、個々の被験者に学力テストの結果を報告する義務はありません。後ほど Plausible Value Technology の説明もいたしますが、NAEP の目的は、個々人の能力推定値を算出することではありません。そうではなく、集団の統計的特性を把握することです。あるエスニシティ・グループで、この辺りの地域に住んでいる、このくらいの階級の人たちの θ (能力推定値) の平均や分散がどうなっているのか、などを知るのが NAEP の目的です。また ETS から学校ごとの学力

ランキングなどを報告する必要はありません。この点をまずお伝えしておく必要があります。

■「何ができるか (パフォーマンス)」を測る

PISA でもそうですが NAEP についても、アセスメントの重点が「何を知っているか (what they know)」よりも「何ができるか (what they can do)」に移ってきています。これだけインターネットなどの情報技術が発達している時代では、知識の有無そのものではなく、いかに情報を取り出して活用して、その結果として何ができるのか、そのパフォーマンスに関心が置かれます。つまり、NAEP もだんだんとパフォーマンス・アセスメントに近づいてきていると言えます。

1つの例がライティングですね。もう1つの例が理科の実験に関するテストです。非常に大変だったのですが、実験キットのようなものを作成し全米の被験者に配付して調査を行ったこともあります。様々な砂を、例えば大粒な砂、小粒な砂、あるいは砂鉄のような鉄分を含んだ砂などを用意し、水やふるいや磁石なども用意して、どのように分離させるか、などの実験を行わせるわけです。大変な労力とお金がかかりますが、そういう実物を使ったアセスメントを行うこともありました。

■教育の説明責任

NAEP の結果は、教育政策やテストのあり方に大きな影響を与えます。きちんと説明責任を果たすことが求められます。また、日本でも当てはまることかもしれませんが、テストの結果が公開されると、どうしても人間の性としてランクの一覧表を作りたくなってしまうものです。州や学区が作らなくても、成績順のランキング表がマスコミなど、どこからともなく出てきます。色々な意味で非常に注目度

スライド 2

NAEP 概略

- 第4、8、そして12学年 / 年齢9、13、そして17歳
- National Assessment: 100,000 students from 2,000 schools
- Main assessment & State assessment: 2年ごと
- 読解、数学、理科、作文、米国史、公民、地理、芸術など
- 学力の国勢調査: 個人の学力を報告する必要はない。
- “What students know and what they can do.”
- Performance Assessment
- 教育の説明責任 Educational accountability
- NAEPは反復継続調査である。
- 変化と連続性: 「変化を測定するときは、測定尺度を変化させてはならない」(Beaton, 1990)

の高い試験です。

■ N A E Pは反復継続調査である

上記のように、色々な側面から影響力を持つ NAEP ですが、その意義は単年度の結果のみに注目せず、継続して長期的な傾向を調査することにあると言えます。

データを毎回、あるいは2年ごと、科目によっては4年ごとに、きちんと積み上げて比較・分析を行う。そのためには、先ほども述べたコモン・ブロックス（共通項目群）を用意するなどの仕掛けを組み込んでおきます。

■ 変化と連続性

このコモン・ブロックスですが、少し変更をかけてしまったがために、長期トレンドの分析結果が若干、怪しくなったようなことが以前にありました。その時の反省に基づき Beaton さんが言った言葉が「変化を測定するときは、測定尺度を変化させてはいけない(Beaton,1990)」です。つまり、コモン・ブロックを使うべきだ。コモン・ブロックを変化させる必要がある場合にも、注意深く共通部分を入れるべきだと言っております。

NAEP の基本設計には長期的に継続調査をする意図がきちんと組み込まれています。信頼されるデータというのは、見識や洞察に基づき、年ごとの比較や州ごとの比較が出来るよう設計されていなければならない。それでこそアセスメントだ、といった思想が踏まえられています。

■ 設問の開発

日本では学習指導要領があり、統一されたカリキュラムで教えられていますが、ご存知のようにアメリカでは全国的に統一されたカリキュラムは存在しません。だからテストを作るときには、まず何をテストするかの枠組みを設定することが、非常に大切です。

■ 主調査における問題作成

しかし、そのフレームワークを作るのがまた大変です。色々な人たちが集まり「教科の枠組み (Subject-Specific frameworks)」つまり、どのような内容を出題すべきなのかを設定します。

NAEP に関連する一連の作業の統括は、NAGB (National Assessment Governing Board、全国調査統括委員会) によってなされます。ワシントン D.C.でミーティングがよく開催されます。私も ETS があるニュージャージーから電車で3時間ほどかけて、よくワシントンでの会議に参加しました。

また、それぞれの教科で専門家による全国委員会が全国調査の実施やガイドラインの運営などを行っています。「何を測るか」については、トップダウンというよりはむしろボトムアップで作っている状況と言えます。

■ Reading を例にとって

リーディングを1つの例としてご紹介します(スライド4, 5)。リーディングのテストで測るべき目的、つまり読解力の定義は、「読者、それから文章、読書体験の内容を含む動的で複雑な相互作用」となります。

先ほども申し上げましたが、特定のカリキュラムに従って学習した内容の習熟度を測定する、というのではなく、むしろパフォーマンス、何が出来るか (can do) を測ろうとしているわけです。現代のアセスメントはこういう形で「パフォーマンス」の測定を重視する傾向にあります。読解の目的、項目のタイプがいくつか設定されています。1つは「文学的経験のための読解」。小説やエッセイの一部など、文学的な表現のものが出題されます。

もう1つは「情報のための読解」。新聞記事など、事実関係を正確に把握するためのテスト項目です。

スライド3

2000年主調査の問題設計 (Item Specification)

- 米国には日本の学習指導要領のような全国的に統一されたカリキュラムはない。
- NAEP設問
 - ETSの指揮の下、教師、教科の専門家、測定論の専門家が協力して教科の枠組み(Subject-Specific frameworks)に基づく設問、課題を開発する。
 - National Assessment Governing Board: NAGB 全国測定統括委員会
 - それぞれの教科について、専門家からなる全国委員会がガイドラインを作成し、それぞれの設問が枠組みに合致していることを確認する。

4

スライド4

読解 Reading 1992 - 2000

- 読者、文章、読書体験の内容を含む動的で複雑な相互作用
- 目的
 - 文学的経験のための読解(reading for literary experience): 小説、短編、詩、劇、エッセイ
 - 情報のための読解(reading for Information): 新聞、雑誌記事、教科書、百科事典、カタログ
 - 課題達成のための読解(reading to perform a task): バスの時刻表、ゲームのルールブック、実験の手順、レシビ、地図

5

最後に、先ほどより何回かご紹介している、実用的な「課題達成のための読解 (パフォーマンスタスク)」。バスの時刻表や、ゲームのルールブック、実験手順など、そのような形の資料がきちんと読めるかを測るテスト項目も用意されています。

このようにお話すると、日本の学力テストと少し違ってのように思いませんか。そうなんです。国としての統一されたカリキュラムがないということもありますが、アセスメントの目的についての考え方の違いが大きいのです。彼らがアセスメントで知ろうとしているのは、一人ひとりの学生が、自分ひとりの力で物を読み、情報を得て、インディペンデントな独立した人間として判断、行動をしてやっていけるか、その力を「読解力」と定義しているわけです。ですから、出題される内容も幅広いものになっているわけです。

■問題作成の手順

問題作成の手順については、細かく資料 (スライド6 ~ 8) に掲載しておりますので、読んでいただければお分かりいただけると思います。1つのテストを開発する際には、このように「リーディング (読解力)」の定義をし、何を測るのかを決めてから取り組みます。これが、この後、色々な団体で項目を作る際の基準になります。なぜ、今日、このように細かな資料をご提示したかということ、1本のテストを開発するには、非常に様々な作業を必要とすることを理解していただきたいためです。

もう1つ強調しておきたいのは、様々な立場からの意見やフィードバックを吸収して行われている、ということです。コンテンツについては教科の専門家、テストの実施にはテスト理論の専門家が関わっています。または1つ1つのコミュニティに出向いて意見を聞くこともあります。

また、学校の先生や親御さんたちに、このテストはこういうものを測ろうとしているのだ、と説明して歩くこともあります。全て時間のかかることではあるのですが、NAEPの性質、出題内容の決定は、それだけ民主的なステップを踏んでやっている、といえます。様々な場面で影響を持つからこそ、多様な専門家グループが働いているわけです。

スライド6

主調査における問題作成の手順 I

- ETSのテスト開発の専門家と様々な教科の専門家が設問を作成し、枠組みの設計に沿って分類する。
- 教科領域のテスト開発の経験をふんだスタッフが必要の面から設問を審査し、適宜修正する。
- テスト開発システムに分類情報とともに設問が蓄えられる。
- テスト開発担当者が設計 (specification) に沿って設問をブロックに集めていく。
- 専門家が、不適切用語、編集上の観点から各ブロックを審査する。
- 設問を個人レベルで施行し、生徒がどの程度理解できるか、ワーディングや形式の面からさらに修正するべき点はないか、といった点について審査する。
- 問題開発委員会 (IDC: Instrument Development Committees) が召集され、問題やブロックが枠組みの設計に合うかどうか、正確に分類されているかどうかを独立に確認する。
- 内容と測定に関する外部の専門化グループが問題の分類について独立に確認する。

7

スライド7

主調査における問題作成の手順 II

- 州測定 (state assessment) プログラムのために、NAEP NETWORKが測定に含まれる全ての問題、ブロック、質問紙を審査する。
- 委員会、NAEP NETWORK、内容や測定の専門家の審査に基づき、テスト開発者が開発バージョンの問題を更新する。
- 全国教育統計センター (NCES: National Center for Education Statistics)、全国測定統括委員会 (NAGB: National Assessment Governing Board)、管理予算局 (OMB: Office of Management and Budget)、データ収集に関する政府の政策に合致しているかどうかを検討する情報管理コンプライアンス部門 (IMCD: Information Management Compliance Division) が施行テストに使われる質問紙や設問を審査して、改訂された試行テスト用バージョンが政府の許可を得られるようにする。
- 試行テストに許可番号を得る。
- 試行用のテスト冊子、質問紙が印刷され、その他の測定用の道具 (例えば、録音テープ、写真、科学実験用具) が作成される。
- ブロックから各設問がテスト開発システムに蓄積される。

8

スライド5

4つのモード 読解スタンス (Reading Stances)

- 最初の理解を形成する (forming an initial understanding)
- 解釈を発展させる (developing an interpretation)
- 自分の考えと反応をまとめていく (engaging in personal reflection and responses)
- 批判的なスタンスを示す (demonstrating critical stance)

読解の領域の測定項目は、全て読解の目的とモードのうちのひとつを反映するように開発されている。

6

スライド8

主調査における問題作成の手順 III

- 試行テストの実施。
- 試行テストを採点、分析する。
- 測定の趣旨に合う設問を選ぶ。
- 教科の専門家が測定にえらばれたブロックを審査する。
- ブロックを不適切用語、編集上の観点などから審査する。
- IDCが召集され、設問やブロックを独立に審査して分類コードを確認する。
- NCES、NAGB、OMB、IMCDが質問紙と設問を審査し、試行テスト用の改訂版が政府の許可を得られるようにする。
- 写真製版用のブロックを印刷用に校正し、正式に承認する。
- その他の測定用の道具 (録音テープ、写真など) の最終バージョンについて、作成を正式に承認する。
- 測定用の小冊子 (booklets) と質問紙を準備し、承認して印刷する。
- ブロックから各設問がテスト開発システムに蓄積される。
- 本番の測定の実施

9

■ プリテストの実施

もう1つ、重要なこととして、プリテスト（試行テスト）の実施についてお話しします。本番のテストの前に、試行テストを行い、ある項目は難度が高すぎるとか、ある項目は後ろに置いておくとか、取り組みが悪そうだななど、各項目や問題冊子構成の検討を行います。

NAEP には全学生が参加するわけではなく、サンプリングテストですので、参加した学生ごとに学力のフィードバックをしてあげられるわけではありません。従って、参加者にとっては、どうしても受験のモチベーションが低くなりやすいのです。そういったことから、試行テストを行い、結果の分析を踏まえた上で注意深く計画を立てなければいけないと考えています。

■ DIF 分析 (Differential Item Functioning Analysis)

これも、試行テストを行う1つの重要な目的です。DIF (Differential Item Functioning) とは、各種のサブグループ、例えば人種やエスニシティ、性別などの違いによって、有利不利が発生する項目はないかどうかの分析を行います。これらも、本テストの前に試行テストで十分に吟味がなされなくてはなりませんので、通常の問題作成の工程として組み込まれています。問題開発において、試行テストというのは非常に重要な役割を担っているといえます。

■ 長期傾向調査の問題設計

NAEP は Long-term Trend を調査することが大きな目的である、と先ほども申し上げました。これは、同一対象の成長を追いかけて調査することを意味するのではなく、ある特定の学年の状態について年度を超えて把握し続けるデザインになっています。

そのため、前年度の試験にリンクを貼って等化を行う必要があります。等化の仕方には様々な方法があるのですが、NAEP では、common population (共通受験者) linking という方法が用いられています。等化の方法は他にも、common item (共通項目) linking などありますが、項目 (item) よりも人間 (population) の数の方が多いですから、それを用いてリンクさせる方が安定しているだろう、ということです。リンキングの技術論についても本は出ていますので、詳しくはお読みいただければ幸いです。

3 心理測定技術の革新

ここからは、NAEP で実現された測定技術の革新についてご紹介します。主なものに3つあります。

1つは Item Response Theory (IRT)、項目反応理論モデルです。NAEP が ETS に移る以前、IRT は使われていませんでした。それがなぜ使われるようになったのか、といいますと、2点目のトピックであるマトリックス・サンプリングを行うためです。受験者に負担のないような実施時間で多くの項目情報を得るために、質問冊子を何種類も作成します。各受験者が異なる冊子に回答しているので、総得点を用いた分析では具合が悪い、ということです。さらに、3点目の Plausible Value Technology というのも、IRT を活用していればこそ、の話です。

このテーマについて、全体的に影響を与えているのは、IRT (項目反応理論) の採用です。それにより、様々なメソドロジーが発生してきたと言えます。しかしながら、古典的なテスト理論、総得点や部分得点、古典的な項目分析などを全く用いないというわけではありません。古典的なテスト理論と IRT の両方の技術を用いながら、試行テストや本テストを分析しています。

3.1 学力テスト項目様式と採点技術

学力テストの項目様式には、多肢選択式 (Multiple-choice question)、作業式 (Performance task)、記述式 (Open-ended question) などがあります。出題の分量全体から見ると、ほとんどが多肢選択式 (Multiple-choice question) で、5肢前後の質問形式が多いです。

作業式 (Performance task) というのは、ライティングが1つの例です。例えば、子どもたちに「宇宙人の円盤が来たらどうするか」などの色々なトピックを与え、それに関して書いてもらうのです。ライティングの課題には叙述型、情報提供型、説得型の3つの種類があり、それらの組合せから学生には必ず最低でも2つのトピックについて作文をしてもらいます。2つ以上受験していないと、共通項目を使つてのリンキングが行えません。そういう意味では技術論からの必要性に対応してテスト設計がなされていると言えます。

記述式 (Open-ended question) については、単語を入れるだけのような単純なものから、ある程度の文章を記入

する発展的なものまであります。

■ 自動採点技術

これだけ大規模の調査を行うために、スコアリングの技術も大幅に革新されました。その顕著な例は、Performance task である作文 (essay) です。初期の段階では、全ての作文について2人の評定者による評価を行っていました。それを、片方を人間、片方を ETS が開発したコンピューターによる自動採点システムのペアで評定させるようになりました。現在では、全て自動化され、人間による作文評定は行っていないそうです。

この自動採点システムのように、NAEP の副産物として技術革新が生まれてきます。実際、e-rater (論文自動採点システム) は NAEP 以外のところでも随分使われています。

Performance task には、作文の他、先ほど少しご紹介しました理科の実験の課題などもあります。面白いものとしては、芸術の調査も始めました。これだけ NAEP が注目を集めると、これもやってほしい、あれもやってほしいと、色々な要請があり、芸術についても、どうしてもやってほしい、と。具体的に何をやるのか、と申しますと演技 (acting) や、歌や楽器の演奏などを行い、ビデオで撮影して、その録画したものを評価する形で行います。

このように、様々な方面からの要請に、どのようにして応えるかというのも、ETS の腕の見せどころです。そして、そのチャレンジに伴ってテスト技術の革新が生まれるという側面があります。そのような条件の下で NAEP が今も進化し続けていると感じます。

また、新しいものを取り入れる場合においても、長期推移を観察するという本来の目的から、常に過去の結果とのリンクを考えつつ新しい方向性を探っていかなければならない。そういう難しいバランスの上に立っているという気がします。

3.2 古典的項目反応分析(Classical Item Analysis)

古典的な項目反応分析 (スライド9) において最も重要なのは無回答反応の分布です。サンプリングで行うアセスメントでは、どうしても回答に対するモチベーションが低いいため無回答が多くなります。特に、後ほどご紹介する BIB デザインにおいてアイテム・ブロックの一番後ろの項目への反応が、どうしても低くなります。その点をどうやってフォローしていくかは重要な課題です。

無回答にも色々ありまして、様々な理由によるスキップ

がある一方で、時間切れと思われるものもあります。また、難しい問題を後ろに置くと受験者がまず取り組まなくなりがちです。いろんな意味で、この無回答の分析というのはきちんとやらなければいけません。

■ 不正解選択肢への反応の分布

不正解選択肢の分析も非常に重要です。成績の高い層での選択率が、正解選択肢よりも高い不正解選択肢がたまにあります。そういうものは、試行テストでも本テストでも念入りに把握しておく必要がありますが、この点については、総合得点をベースにした古典的手法の方が、一目瞭然で比較的分りやすいようです。

他には、学力の程度や集団による項目反応の違い、選択肢選択状況の分析などもあります。双列相関係数 (biserial correlation coefficient) も使います。古典的項目分析の方が分析時に有効である点もいくつかあるので、IRT を取り入れてはいても、古典的分析は必ず行っています。

■ 項目の差異分析 (DIF)

DIF (Differential Item Functioning) の分析も行います。DIF というのは、ある項目についてサブグループによる反応の違いが生じるかどうかを調べるものです。同じ能力であっても、たまたま所属しているサブグループによって反応が異なるような出題は困ります。日本でもこれから、このような観点が大切になってきます。

例えば、数学の質問に野球の例を入れると、女性に不利になりやすいとか、バスケットボールだと黒人のほうが有利だけれども、レガッタなどは白人の上流階級で有利であるなどのケースがあります。このように、階級や人種などの違いで、項目反応に影響の出る項目をできるだけ排

スライド9

古典的項目反応分析

- 二値項目反応(dichotomous item responses)の分析
 - 被験者の問題回答行動の分析
 - 無回答反応の分布
 - 不正解選択肢への反応の分布
 - 学力の程度や下集団別による項目反応の分布の違い
 - 項目識別度としての双列相関係数(biserial correlation coefficient)
 - 下位集団別の選択肢ごとの反応度数(response frequency)などの分割表(contingency table)

除するようにします。

しかしながら、「バイアス」があると断定するためには、統計量だけでは不十分です。専門家や、事情をよく知っている人たちの討議を経て、統計的に現れたバイアスの実態について、説明がつけられる必要がある。

手法としては Mantel-Haenszel の統計法を使います。これは、古典的手法です。

私自身も、NAEP ではありませんが、DIF の検証で呼び出されてヒアリングを受けたことがあります。当時の ETS には東洋人が少なかったので、よく呼ばれていました。1 つ例を出すと、自転車素材にした問題で呼ばれたことがあります。アメリカでは自転車はレジャーですが、中国ではレジャーではなく通勤ですね。そうすると、自転車素材にした問題で技術的な質問をすると回答の特性に差が発生します。言われてみればそうか、ということなのですが、その第一発見の方法として項目分析の統計量が活用されている、というイメージです。

■distractor 分析

先ほどお話しました distractor(錯乱肢)分析、不正解選択肢の代替案の検討なども重要です。

■多値項目反応 (polytomous item response) の分析

NAEP は高度の思考過程を計ろうということで、パフォーマンス・アセスメントに近くなっている、という話を先ほどいたしました。最近の傾向として、「何を知っているか」ではなく「何が出来るか」を測ろうと、高度に知的な能力が必要になる項目をなるべく多く取り入れようとしています。

そこで、当然ながら多肢選択形式よりも、記述式 (Open-ended question) の項目が多くなります。単なる Open-ended question というのは単純に正解、不正解が決められるものですが、Extended open-ended question (長文記述式問題) になると、100%ではないがほぼ正解、とか少しだけ点数をあげたい、とか部分的な正解を認める partial credit model が必要になります。

日本においてもそうだと思いますが、アメリカにおいても多肢選択方式について再考すべきという意見があり、constructed response item (記述回答式項目) の使用が増えています。作業式の項目もそれに含まれますが、単に知識の有無を聞くのではなく何が分っていて何が分からないのかを自由な解答から測定する項目です。どのような能力を測るべきか、という新しい能力観のようなものが背景に働いています。

一番典型的な例としては Essay type の課題が挙げられます。私は、この Essay type のアセスメントに比較的深く関与していました。評定者を集めて、最初はルーブリックを作るところから行うわけですが、その採点基準作成や一貫性の確保も大変です。

このように、多値反応の項目というのは増えていますが、必然的に無回答が多くなりやすい。それをどうするか、ブロック内の問題の配列を含めて課題になっています。

3.3 項目反応理論 (Item Response Theory) モデル

項目反応モデルにおいて、項目に関する未知の情報が item parameter であり、その推定が calibration です。

■標本加重の再計算

NAEP の目的は、サブグループごとの能力値の分布を正確に測ることにあります。例えば、黒人やアメリカン・インディアン等の学力についての情報が重要なのですが、彼らを実際の人口比に応じてサンプリングするとほんの少しの人数しかサンプルが得られないことになります。しかし、サンプルが少なすぎでは十分な情報が得られません。

そこで、どうするかというと、まず人口比で計算したよりも多めの人数をサンプリングします。そして item parameter を推定するときには、そのサンプルの比を低く抑えるのです。つまり、over representation をして、その後、sampling weight を低めて計算します。これを、標本加重の再計算といいます。

また、サンプル調査をした時点と実際の NAEP の本調査の時点には、間隔がありますから、その間に人口移動があり、分布が変わっているかもしれません。ですから、sampling weight は、本調査でもう一度再計算されます。このように統計の知識をフル活用して行っています。一度、抽出をした後になお階層化を行うので、事後階層別 (poststratification) と呼ばれています。

■IRTモデルの決定

IRT のモデルは何を使用するかというと、項目の回答様式で異なります。Yes / No のような 2 値反応の項目については、three-parameter logistic model を使います。3 つのパラメータとは、項目識別力 (item discrimination)、項目困難度 (item difficulty)、当て推量パラメータ (guessing parameter) です。サイコメトリシャンとしては、この当て推量パラメータの推定に最も苦労しました。

多値反応の項目については、guessing を考慮しなくてよいわけなので generalized partial credit model という

ものを使います。2つのモデルを使い分けています。

解析については、ETS / BILOG PARSCALE というプログラムを使いました。IRT のプログラムにはいくつかありますが、そのうち BILOG - MG と PARSCALE は、どちらも私が書いたもので市販されています。その2つをまとめて、ETS で新しく書いたのが ETS / BILOG PARSCALE です。

■Logistic Item Response Model (2PL)

Logistic Item Response Model についてご説明します (スライド 10)。傾きが item discrimination または slope parameter, y 軸 (縦軸) の確率 0.5 と曲線が交差するところの x 軸 (横軸) の値を item difficulty あるいは thresholds と呼んでいます。x 軸は能力 θ で、これが高くなるほど正解確率は上がる、というモデルです。

このモデルの式の中の a や b が推定されたパラメータです (スライド 11)。その推定値が確からしいとかおかしいとかの感覚が必要になるわけですが、その背後にはある程度の心理測定論的な知識が必要とされます。

b (項目困難度) の概念は、理解しやすいと思います。b が高ければ高いほど項目は難しい、というようになります。スライド 12 において、 $P_1(\theta)$ と $P_2(\theta)$ では b の値のみが異なります。同じ形で横に平行に移動します。もちろん、項目 1 の方が難しく、項目 2 の方が易しい、ということです。例えば、横軸 (能力値 θ) の値が 1.0 の人であれば、項目 1 に正解する確率は 10% 程度ですが、項目 2 であれば、ほぼ 100% である。より正確に言うと、1 人の人の正解する確率ということではなくて 100 人の何割が正解する確率、という説明の仕方の違いはありますが、おおよそそのような感じです。スライド 13 の $P_1(\theta)$ と $P_2(\theta)$ に

おいては、b (項目困難度) の値は同じですが、a (項目識別力) の値が異なります。受験者の能力の違いをより正確に区別して測ることができる項目は、実線の項目 1 ($P_1(\theta)$) の方です。

スライド 11

Normit or Logit

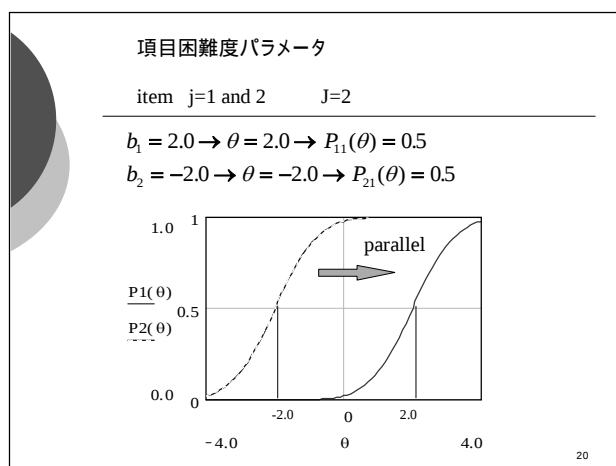
$$Z_j(\theta) = a_j(\theta - b_j)$$

a_j item discrimination
discriminating power
slope parameter
項目困難度

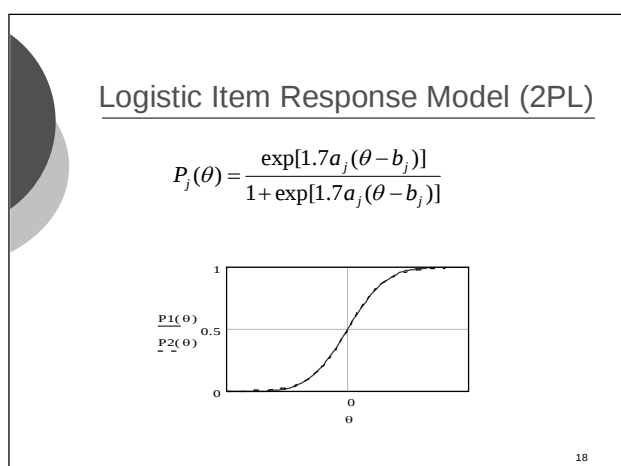
b_j difficulty parameter
threshold parameter
item location parameter
項目識別度

19

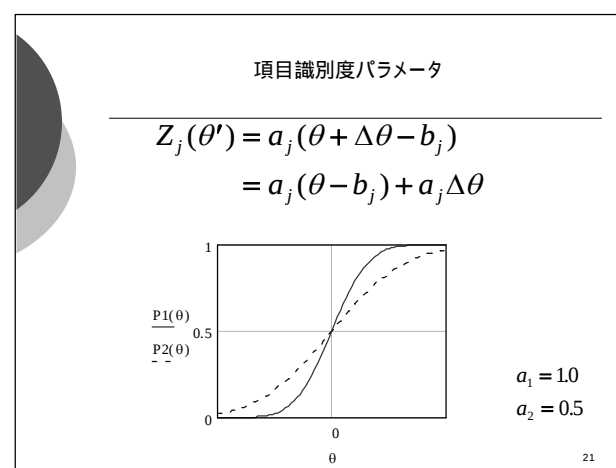
スライド 12



スライド 10



スライド 13



■ 3 Parameter Logistic Model (3PL)

スライド 14 では、先ほど触れました当て推量パラメータ (g) を入れた、3 パラメータロジスティックモデルを紹介しています。

■ Polytomous (多値型) Item Response Models

多値型項目についても IRT を使うかどうかについては、初期の頃に色々ともめました。私も、ETS が最初に IRT を使うと言い出した頃、Journal of Educational Measurement に Extended open-ended question にはどのモデルを用いるべきかについての記事を執筆しました。当時は、Graded Response Model を使おうかなんていう話もありました。その頃出版した私達の論文を見ていただければと思います（スライド 15）。

結論としては、どちらでもよいと思うのですが、個人的な考えからすると Graded Response Model の方が推定を行いやすく、よいのではないかと思います。しかし、当時は Partial Credit Model が、色々な意味で人気が出てきているときでした。ネーミングも多値型に合っている感じがするし、どちらでもいいなら、こちらにしよう、ということで現在 NAEP では Partial Credit Model を使っています。

スライド 14

Logistic Item Response Model (3PL)

$$P_j(\theta) = g_j + (1 - g_j) \frac{\exp[1.7a_j(\theta - b_j)]}{1 + \exp[1.7a_j(\theta - b_j)]}$$

where

当て推量のパラメータ g

スライド 15

Polytomous Item Response Models

- Graded Response Model (Samejima, 1962, 1972; Muraki, 1990)
- Partial Credit Model (Masters, 1982)
 - Generalized Partial Credit Model (Muraki, 1992, 1993)
 - Rating Scale Model (Andrich, 1978)
- Nominal Response Model (Bock, 1972)

■ Generalized Partial Credit Model

中でも特に a パラメータ (discriminator) がついている Generalized Partial Credit Model (スライド 16~18) を使っています。本当は、私の恩師でもある、Bock 先生の提唱する Nominal Response Model (名義反応モデル) もデストラクター (錯乱肢) 解析に使えばよいと思うのですが、NAEP では使われていません。Multiple-choice では 3 Parameter Logistic Model, Polytomous の場合は Generalized Partial Credit Model を使います。先ほど BILOG と PARSCALE の混合した形のものを使っていると申しましたが、要するに BILOG で Multiple-choice を、PARSCALE で Partial Credit Model を分析しています。

スライド 16

Generalized Partial Credit Model

$$P_{jk}(\theta) = \frac{\exp[\sum_{v=0}^k Da_j(\theta - b_j + d_v)]}{\sum_{c=0}^{K_j} \exp[\sum_{v=0}^c Da_j(\theta - b_j + d_v)]}$$

$$= \frac{\exp[\sum_{v=0}^k Z_{jv}(\theta)]}{\sum_{c=0}^{K_j} \exp[\sum_{v=0}^c Z_{jc}(\theta)]}$$

$$= \frac{\exp[Z_{jk}^*(\theta)]}{\sum_{c=0}^{K_j} \exp[Z_{jc}^*(\theta)]}$$

25

スライド 17

$$Z_{jk}^*(\theta) = \sum_{v=0}^k Z_{jv}(\theta)$$

$$= \sum_{v=0}^k Da_j(\theta - b_j + d_v)$$

$$= Da_j[k(\theta - b_j) + \sum_{v=0}^k d_v]$$

$$= Da_j[T_k(\theta - b_j) + K_k]$$

T_k Scoring Coefficient
 0 1 2 ...

K_k Category Coefficient

26

スライド 18

$$P_{jk}(\theta) = \frac{\exp[\sum_{v=0}^k Z_{jv}(\theta)]}{\sum_{c=0}^{K_j} \exp[\sum_{v=0}^c Z_{jv}(\theta)]}$$

$$= \frac{\exp[Z_{j0}(\theta)] \exp[\sum_{v=1}^k Z_{jv}(\theta)]}{\exp[Z_{j0}(\theta)] + \sum_{c=1}^{K_j} \exp[Z_{j0}(\theta) + \sum_{v=1}^c Z_{jv}(\theta)]}$$

$$= \frac{1 + \exp[\sum_{v=1}^k Z_{jv}(\theta)]}{1 + \sum_{c=1}^{K_j} \exp[\sum_{v=1}^c Z_{jv}(\theta)]}$$

27

多値型項目の応答特性曲線を示しました(スライド 19)。Dichotomous (2 値型) のモデルでは、線が 2 本しかなかった(スライド 20)わけですが、このモデルの場合は 3 カテゴリーを考えます。全くできない確率の曲線 ($P_{j0}(\theta)$)、きちんとできる確率の曲線 ($P_{j2}(\theta)$)、中間に少しは回答できる確率の曲線 ($P_{j1}(\theta)$) があります。

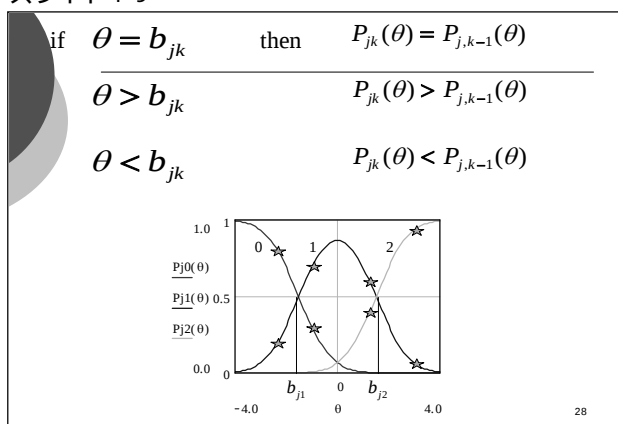
θ がだんだんと右にずれて、能力が高くなってくると、全く出来ない確率が下がって、少しは出来る確率が高まります。 b_{j1} あたりで、全く出来ない人と少しは出来る人の割合がちょうど半々になります。

■ パラメータの推定 (Parameter Estimation)

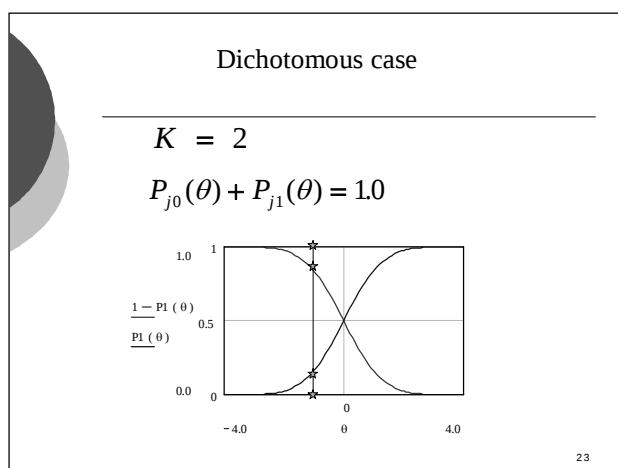
Parameter の推定にも、色んな方法があります。この点においても、NAEP が世界のサイコメトリック界に影響を与えたのは、Bock による周辺最尤推定法(Marginal maximum likelihood method)です。

これは本当は Maximum marginal likelihood らしいのですが、まあどちらも MML と略せます。NAEP はこれで、パラメータの推定を行います。当時は、同時最尤推定法(Joint Maximum Likelihood) という Fredrick Lord

スライド 19



スライド 20



が提唱した推定方法があったのですが、MML がそれにとって代わるようになりました。

MML とは、どういうものかといいますと、潜在特性 (latent trait) この場合、測定対象となる学力 (能力) を random component とします。つまり、事前確率分布を仮定します。事前確率分布があるということ、それは Bayesian method ということです。Bayesian であれば、今はモンテカルロ法などの computer intensive な方法があります。これは、コンピューターがこれだけ高速になったからできるものです。NAEP では実用化されていませんが、研究対象にはなっています。

何を言いたいかというと、こういう方法論的な部分と、コンピューター関係の技術論的な部分、さらに教育論的な思想がうまく絡み合って NAEP が発展してきているということです。

「“全米学力テスト”なのだから、全員に同じ問題を解かせればいい」と言ってしまえば、それでお終いです。しかしそれでは学力測定として非常に限られたものになってしまいます。そこで、項目や受験者をサンプリングするのであれば、どういう方法で行うのか、各サブグループをどのように正確に描写できるのか、などの色々な問題に対応しようとしてアメリカにおける測定は発達してきたのです。

■ 潜在特性の推定 (Latent Trait Estimation)

IRT における個人の能力の推定 (point estimate) には MLE(Maximum likelihood estimation) や WMLE(Weighted maximum likelihood estimation), EAP(Expectation a posteriori) や MAP(Maximum a posteriori) など、様々な方法があります。しかしながら、NAEP で潜在特性の推定に用いる Plausible Value (推算値)は、それらと考え方がかなり異なるものです。

NAEP においては、個々の受験者の能力推定を行う必要はありません。NAEP の目的は、全米の児童生徒全体の学力分布、またはエスニシティなどによるサブグループの学力分布をできるだけ正確に推定することです。

MLE などの個々人の能力を推定する方法で求めた能力値を集計すると、例えば白人とアメリカン・インディアンのように集団の大きさに極端な差がある場合、個々人の推定値から集団の分布を推定する方法ではバイアスが大きくなってしまい、正確な分散の比較になりません。

そこで、Plausible Value を用います。Plausible Value は、先ほどお話しした Bayesian のフレームワークです。

Bayesian のフレームワークを可能にした発端は、先ほどの項目パラメータ推定に使われる Maximum marginal likelihood method で、特に EM(Expectation Maximization) アルゴリズムを使用しています。

現在、統計学関係の方は皆さん EM アルゴリズムを知っていらっしゃると思うのですが、その当時は画期的なものでした。繰り返しになりますが、NAEP というのは、このような統計学あるいは教育測定学における比較的先端の技術を導入しながら発展してきたと言えます。

■等化 (Equating and Linking)

等化についても、色々な方法が考えられています。IRT 等化法は効果的です。スライド 21 にご紹介した文献に詳しく書きましたので読んでいただければと思います。

IRT 等化法には幾つも方法がありますが、NAEP では安定性が高いという理由で Common Population (共通受験者) 法を採用しています。なぜ、Common Item (共通受験者) 法を採用しないかについては、項目よりも受験者の方が圧倒的に情報が多いためだと、先ほども述べました。

3.4 マトリックス標本抽出方法と B I B デザイン

先ほど出てまいりましたマトリックス標本抽出法について、ここで詳しくご説明します。マトリックスですから、縦と横の2方向があります。縦のサンプル(標本)とは、「被験者」です。NAEPは、全数調査ではありません。まず地域を選んで、学校を選んで、その学校の中で被験者をサンプルするという流れになっています。

一方、横のサンプルとは「テスト項目」です。要するに、いくつかの項目の集合で「ブロック」を作ります。そして、ブロックAとブロックBを組み合わせるとか、ブロック

BとブロックCを組み合わせるなどしながら1冊のブックレット(冊子)を作ります。その際、共通のブロックに入っている項目(common block items)を活用しながら、体系的にリンクを貼っていきます。

なぜ、このようなサンプリングを行うのでしょうか。NAEPは「調査」ですから、できるだけ多くのことを調べたいわけです。様々な内容のたくさんの項目を使って色々な分野、領域についての調査をしたい。しかしながら、1人の受験者から全ての情報を得ようとする試みは、試験時間が何十時間もかかってしまいます。標本に該当した学生は、呼び出されて何十時間もただで項目をやりとげなければならない。授業中に呼び出すわけですから、そんなに長い時間、学生を拘束することはできません。

そこで、質問できる項目の数を増やすために、複数の冊子を用意して、その中のブロックを組み合わせながら調べます。また、できるだけ当該被験者のレベルに合った項目で構成される Booklet を回答してもらるようにコントロールします。回答時間や提示の順番にも配慮がなされます。

このことを可能にしたのが IRT です。なぜなら IRT というのは回答者と1つ1つの項目のインタラクションを記述するモデルですから、項目のパラメータが分れば、別の項目であっても、使われた冊子が違っていても、きちんと θ が推定できて、同じスケールに乗せることができるのです。ですから、このマトリックス・サンプリングの考え方は、IRT の手法がなければ実現が難しかったということになります。

■つり合い型不完備ブロック計画 (BIB design)

つり合い型不完備ブロック (Balanced incomplete blocks design) と書いてあります。何が incomplete なのか、例(図1)を挙げて説明します。

これは、1990年の数学のMain NAEP で用いた、Balanced incomplete blocks design です。項目のブロッ

図1 B I B デザインの1例 - 1990年数学メインNAEP

小冊子のバージョン	配列1のブロック	配列2のブロック	配列3のブロック
1	A	B	D
2	B	C	E
3	C	D	F
4	D	E	G
5	E	F	A
6	F	G	B
7	G	A	C

スライド 21

Equating and Linking 等化

Muraki, E., Hombo, C. M. & Lee, Y. (1999). Equating and linking of performance assessments. *Applied Psychological Measurement*, (in print).

- Classical Equating Methodologies
 - Linear
 - Equipercentile
- IRT Equating Methodologies
 - True score
 - Common item - TCC and ICC's
 - Common population - example: NAEP
 - Multiple-group IRT model

クは、AからGまであります。1つの冊子には、3つのブロックが配列されており、その3ブロックの組み合わせ方で7とおりの冊子を用意しています。実現していない組み合わせもあるので Balanced “incomplete” blocks design (BIB) と呼ばれています。不完全といいますが、そこはリンクを貼っているのです。これがもし全ての組合せで調査するとなれば Balanced “complete” design, BCD と呼ばれるでしょう。

被験者によって異なる組み合わせのブロックを与えることで、測れるコンテンツ、ドメインを拡大していく、という考え方です。「被験者である児童生徒が NAEP テストに割かねばならない時間を最小限に抑えながら、しかし広範囲な学力を正確に測ること」がこれらのテストデザインの目的です。Pophan という人が初期の理論を構築しました。

■ 層化多段階抽出法 (Stratified Multi-stage Sampling)

層化 (stratified) 抽出法とは、母集団から直接、単純無作為に標本を抽出するのではなく、予め設定されたグループから、それぞれ必要な標本を必要数、無作為に抽出する方法です。これを、地域、学校、学校の中の生徒、と全てのレベルで行うので Multi-stage となります。

地域については、アメリカを4領域に分けます (スライド22)。また、アメリカは大都市圏 (Metropolitan statistical area, MSA) と、非都市圏 (Non-MSA) で、はっきりとした違いがあります。さらに、学校の性格として公立と私立の違いが大切になります。先ほど述べたオーバー・サンプリングを行ったところは、後からサンプリング・ウェイトで調整をつけるなど、色々なところでコントロールを行っています。

■ SD / LEP

NAEP がこれだけの注目を浴びようになると、社会的な責任が出て参ります。その1つに、SD や LEP の学生も測定に加えるべきだという動きがあります。

SD とは、Students with disability. Disability には、認知的なものも、そうでないものも含まれます。LEP とは Students with limited English proficiency で、移民などが該当します。そういう人たちを無視してはならないということで NAEP にも参加してもらっています。私が ETS にいた当時は、こういう人たちの受験が容易になるような施策はとられていましたが、実際の分析には加えませんでした。今は、どのようになっているのか分かりません。

日本では考えられにくい発想かもしれませんが、アメリカ

力では「試験を受ける権利」が主張されます。つまり、試験を受けるということは、測定の結果に基づいて何かしてもらう権利だということです。日本のように、学力テストをやればみんなが勉強して学力が上がるだろう、というような発想ではありません。むしろ被験者の立場から見ると自分達のことをよく知ってもらうための権利です。学力テストの結果に基づいて色々な政策が決められる際に、テストを受けていない人の情報は反映されないからです。当然、disability のある生徒の学力スコアは低いと思いますが、その低いという情報をきちんと分析してもらう権利が必要だ、ということです。そういう意味では「テストを受ける権利」というのは重要だと感じます。

■ BIB デザインとテスト冊子

BIB デザインを使うメリットとしては、一人当たりの生徒の検査時間が短くても広範囲の情報が得られることその他、隣り合った生徒同士の冊子内容が異なるので不正行為が生じにくい、という利点も挙げられます (スライド23)。

その他に、学生や学校の背景情報 (Background Variable) を質問紙 (Questionnaire) で集めるのにも有利です。例えば、テレビを何時間ぐらい見ますかとか、宿題にどのくらい時間かけますか、など政策に直接提言するのら

スライド22

層化多段階抽出法
(Stratified Multi-stage Sampling)

- ユニバース Universe
 - 地域 4 regions
 - 北東地区、南東地区、中央部、西部
 - 都市部 (metropolitan statistical area, MSA) & 非都市部 (Non-MSA)
 - 第一次抽出単位 (Primary sampling unit, PSU)
 - Oversampling for minority groups
 - 1996 Main 94 PSU, Long Term Trend 52 PSU
 - 第二次抽出単位 (Secondary sampling unit, PSU)
 - 学校・公立 (public) と非公立 (non-public) school
 - Oversampling for non-public schools and schools for minority groups
 - 最終抽出単位 児童生徒 Oversampling
 - Sampling Weights で過剰抽出を調整
 - Poststratification: 標本加重の再計算

スライド23

BIB デザインとテスト小冊子

- ブロック Block を組み合わせてテスト小冊子 Test Booklets
 - 生徒間の不正行為を防ぐ。
 - 一人当たりの問題数が限られても、全体では広範な学力内容についての情報が得られる。
 - 測定する教科内容の範囲を犠牲にせずに試験時間の負担を軽減できる。
- Focused Spiral BIB Design
- Subject Specific Background Questionnaire, Demographic Background

3.5 Plausible Value Technology

能力値 θ ,それから Background variable , Response pattern などから生徒一人ひとりの事後確率分布(the posterior distribution) を予測します。そして、地域別やサブグループ別の分布を推定します。

生徒一人ひとりの θ は計算していません。この方法の利点は、児童生徒全体、あるいはその下位集団の分布について、より正確な予想ができることです。つまり、point estimate したときの一つひとつの θ に含まれる不確かさの影響が軽減されるわけです（スライド 24, 25）。

また、マトリックス・サンプリングによって生じる受験者ごとの測定誤差分布が大きく異なるという課題にも対応します。

スライド 24

PV: なぜ個人の θ (Point Estimate) の集計ではいけないのか？

- 個人の受験者の能力推定値に含まれる誤差は無視できない大きさとなり、この能力推定値を用いて推定される母集団の能力値分布は、真の能力値分布を反映したものとならない可能性が大いにある。
- 受験者ごとに項目数や形式、内容の異なるテスト冊子が実施される場合、測定誤差分布が異なり、母集団の能力値分布を正しく推定することができなくなる。

スライド 25

PV

Rubin's Multiple Imputation

$$t^*(x, y) = E[t(\theta, Y) | x, y] = \int t(\theta, y) p(\theta | x, y) d\theta$$

IRT尺度下におけるPVの算出

$$p(\theta_r | x_r, y_r, \Gamma, \Sigma) \propto p(x_r | \theta_r, y_r, \Gamma, \Sigma) \times p(\theta_r | y_r, \Gamma, \Sigma) \\ = p(x_r | \theta_r) \times p(\theta_r | y_r, \Gamma, \Sigma)$$

4 質疑応答

■ リンキングについて

—— リンキングについて教えてください。
common item（共通項目）ではなく common population

（共通被験者）で行っている、ということでしたが。

村木：例えば、同一の被験者が一部の共通項目を含む、異なるブックレットに取り組んだとしたら、ブックレットが違っていても共通項目については同じ特性分布になるだろう、という仮定が入っています。だから、common population です。実際には、この仮定にはかなり危険な側面があります。

■ 長期トレンドについて

—— Long-term Trend NAEP も実施されていますが、どちらかというと、そのような長期トレンドを検証するよりも、1 回ごとのテストにおける人種や地域による違いを検証することに重きが置かれているのでしょうか。

村木：もちろんアセスメントですから、Long-term Trend は大事です。それに、集団ごとの傾向を見ることも大事です。私が申し上げたかったのは、同じ集団の推移を追いかける観点、4 年生の時に測定した同じ生徒の 8 年生での成績の変化の検証をするという観点には力を入れていない、ということです。そのような成長を直接測る縦断的調査が主要な目的ではないからです。

■ NAEP が教育政策に与える影響について

—— NAEP が持つ影響力について、お伺いします。
アメリカ国内で非常に影響力を持っているということでしたが、具体的には、何にどのような形でインパクトを持つのでしょうか。というのも、アメリカには統一のナショナル・カリキュラムが無いわけですし、これだけパフォーマンス・アセスメントなども工夫して丁寧に収集した情報をどのように教育政策に反映しているのでしょうか。

村木：ETS 自体は教育政策について、ああすればいい、こうすればいい、というような提言は行いません。レポートを見ていただければ分りますが、単純に事実を羅列しています。レポートを自動生成するプログラムがあって、ちょっと見ると無味乾燥なものが出てきます。ある意味、意図的に価値判断が入り込まないようにしています。われわれは、事実の情報を提供します。そこから政策への示唆を読み取って判断するのは、受け手に委ねています。

学区の当事者は、このデータを使って学校ごとの順位などをつけたがります。これは新聞等を通じて保護者にも公開され、大きな影響を持ちます。それから、研究者に対しては教育に関する様々な条件、変数のデータを提供します。例えば、テレビの視聴時間と成績の関係など研究したければ NAEP のデータは公開されていますから使ってもらうことができます。また、政策の当事者が何か判断しようと

した際に、1960年代からの積み上げがあつてかつ信頼の置けるデータはNAEPしかありません。データに基づいて政策判断を行わねばならないときに非常に重要なソースになります。ただ、繰り返しになりますがETS自体は、NAEPの結果に基づいて何かすべきだ、というような政策提言めいたことは言いません。

—— アイテムを開発する過程で議論される内容、つまり、「これからの時代には、こういう能力が必要なのではないか、だからこんな問題を作ろう」というようなことは、非常に強いメッセージ性を持っていると思います。非常に工夫して項目を開発していても、「こんな能力を測っていくべきだ」とか「学校で育てるべきだ」というような点をレポートで大きく取り上げることはないのですか。

村木：私達は、同時代人ですから、同時代人というのは意識するしないに関わらず、何らかのトレンドを共有しています。もちろん、哲学的・思想的なもの、教育観や人間の能力に対する考え方など。それからコンピュータ・テクノロジーの影響もありますよね。そういった影響を同じ時代の中で受け止めているから、これからの時代に必要な能力について、測定のフレームワークなどを考えていても意図せず自然とコンセンサスが持ててくる。それは、強力なオピニオンリーダーがこうしろ、ああしろと押し付けて得られるものではなく、自然とたどりつくものだと思います。

—— 先ほどの質問者の発言にあった、アメリカには統一のナショナル・カリキュラムが無い中で、この調査結果をどのように活用しているのか、という点についてコメントします。

各州で、勝手に色々なことを行っている中に、統一のテストを行ってみるということは、教育実験とその結果検証が自然と出来ている、と言えるだろうと思います。日本のカリキュラムは全国統一で、カリキュラムの変更も全国一斉に行われます。ですから、対照実験のようにして検証する術を持たない。そういう意味で、アメリカにおいて州ごとに違うプログラムが同時平行で進行していながら、ある時点で共通の指標でその成果を測っているというのは大変興味深い。測定内容にも、パフォーマンス・アセスメントなど、新しい観点が入って進歩的ですよね。

項目形式について質問しますが、多肢選択式の見直し、というか多肢選択ではない方向へ向かっていっているのでしょうか。

村木：NAEPに限らず、その傾向はありますね。

—— NAEPの内容は非常に進歩的であると思いますが、その一方で伝統的な標準学力テストも実施されてい

ますよね。例えば、ある州が伝統的なテストでは比較的よい成績なのに、NAEPで測るような高度なスキルやパフォーマンス・アセスメントでは、そうでもない、というように、多角的な比較が可能になると評価も深まるように思うのですが。

村木：先ほど、学力テストのデータが無いと言いましたが、SATやACTのデータはあります。その比較なども行われています。しかし、色々な学力調査があるものの、NAEPが実際には柱になっていますし、それぞれ違う意図のテストですから、比較には注意が必要です。また、地方によっては、進化論の取り扱いなど、出題に関する意見も異なり同一というわけにはいきません。

—— NAEPの性格について整理をするために少しコメントをします。NAEPの性格、役割というのは比較的、明確になっています。私は政治学などの分野は明るくないので、それこそ社会学や政治学の方にも少し研究をいただければよいと思っておりますが。

NAEPでは、それぞれの州や全国の学力状況についてのレポートがETSから出されます。ETSの役割は、あくまで正確な情報を提供することであって、それをどのように使うかは、その州政府や議会で話し合われるべきことなのです。NCLB法が成立してから、各州がさまざまなことを行っています。NAEPのデータも各州や学校で、それぞれの政策を考えるベースになります。ETS側から、こういうことをやったらよい、というような提案はしません。正確な事実の情報を提供することが重要です。例えばある州である教育政策を立てたときに、そのベースになっているデータが怪しいのではないかと、となると大変なことなので、そういう不安が生じないようにきちんとした調査を行いましょう、という姿勢です。

また、先ほど話題になった新しい時代に対応するものとして、どのようなアイテムを入れればよいかという議論はNAGBのフレームワークの中で考えられるものです。

■測定内容のフレームワーク作成

—— 今いただいたコメントで、ずいぶん整理できましたが、新たな疑問もわいてきました。新しい時代に対応するような測定項目であればあるほど、それが出されたときに、何を測ろうとしているのか、現場レベルで理解されるのが難しいのではないかと思います。教育政策にどのように生かすかを議会の議員が議論することで現場レベルでの理解が広がっていくとは考えにくい。現場の先生が所属する色々な団体が教科ごとにあると思うのですが、そこ

で NAEP の測定内容について理解を深めるために調査結果を解釈するような取り組みがなされているか、についてご存知であれば教えてください。結果について独自の解釈を行ったり、指導法の提案をしたり、ということがなされているのかどうか。

村木：思考方法が逆だと思います。アメリカでは、まず先に「このような能力を測る」というフレームワークを現場のコンセンサスも得ながらきちっと描くのです。測りたいもののコンセンサスが先にあり、その後、ではそれを計るためのテスト項目は誰が作るのかという別の専門家です。ですから、「議論しているうちに大事なテスト項目が出来ました。やってみました」のような形で、唐突に提示されるわけではないのです。ですから、現場を突然に驚かすような項目は少ないと思いますよ。

——— 少し補足のコメントをいたします。Main NAEP の中には、Long-term trend NAEP で質問するような読解や数学とは違って、その時々に関心に応じてその年だけ作成する項目、というのがあります。それを見ると非常に面白い。例えば、どういうものがあるかということ1980 年代にパソコンが普及して、多くの学校で使われるようになった際、子ども達がコンピューターをどの程度、どのように使っているかの調査を取り入れました。また、それよりも少し前、1970 年代ぐらいでは、テレビの普及とともにコマーシャル・メッセージが浸透して衝動的消費が問題になってきた。そういうときには、消費者教育的な内容を出題したり、また金融リスクや消費者金融のようなものを取り入れたり、時代のニーズやリスクに合わせた出題がなされています。これは非常に大事なことだと思います。その結果を基に、今の子ども達はこうだからと考え、時代に必要な教育政策を検討し、それを測るにはどうしたらいいか、ということを行っているようです。

村木：Long-term trend NAEP と Main NAEP には、かわりを持たせています。

——— 先ほど、フレームワークの設定には、例えば数学で言うと NCTM (National Council of Teachers of Mathematics) のような教師団体が、それぞれ関わっているというお話でした。芸術を含む各教科で、それが行われているとすると、非常に大規模になるし、かつ4 年おきに調査を行うことを考えると準備段階も含めて意見がまとまらずに紛糾したり混乱したり、ということはないのでしょうか。誰かが、強力なリーダーシップを発揮してまとめる、というようなことがあるのでしょうか。

村木：私は測定の専門家なので、細かいことは、あまり

よく知りませんが、フレームワークの作成場面においては、コンセンサスを非常に大事にしていると理解しています。現場の先生方はもちろん、保護者を招いての公聴会のようなことも行っていると聞いています。もし、強引にまとめあげようとリーダー的な方が強圧的なことをやったとしたら、それ自体が続かなくなるのではないのでしょうか。大事なのは、コンセンサスの形成、その素地としてはパブリケーション、PR ですね、大きな混乱が生じていないから、やっていけているのだと理解しています。

■実施の負担感について

——— イギリスの事例なのですが、1990 年代に同様に国の学力調査にパフォーマンス・アセスメントのようなものを導入しようとした時に、実施のコストを理由に教師達のボイコットなどが起こって、一旦、実施が取りやめになったということがあります。その後、もう少しコストを下げた別の形式で導入したようなのですが、イギリスでは上手く行かなくてアメリカで混乱なく続いている理由について、先生のご見解があれば教えていただきたいです。

村木：それだけ、結果として提供される情報量が豊富であるということもありますしマトリックス・サンプリングなどで実施の負荷をできるだけ軽減するよう工夫してきたこともあります。それから、先ほど申し上げたパブリックリレーションズですね。

しかし、ETS にも決して全校が諸手を上げて協力してくれているわけではなくて、校長先生から時間が無いとか、その他の理由で拒否されることもあります。拒否された場合には、同じような条件の代替の学校に依頼できるよう情報を持っていないといけません。断られることが無い、というのではなく、断られたときにどうするのか、の対策をいつも持っている、ということでしょうか。

引用・参考文献

- 1) 「全米学力調査 (NAEP) の研究」 全米学力調査研究会 (代表 荒井克弘) 2005 年 7 月

以上