課題名　全米学力調査（NAEP）の研究

研究代表者名　　荒　井　克　弘（教育政策科学講座教授；教育計画論）

研究組織（研究分担者等）

村　木　英　治（大学院教育情報学研究部・教育部）

倉　元　直　樹（高等教育開発推進センター高等教育開発部
入試開発室 / 大学院教育情報学教育部）

研究目的と方法

本研究では，1990 年代を中心に，全米学力調査 NAEP （the National Assessment of Educational Progress）で行われてきた調査の方法論とその成果について調査を行い，わが国の今後の教育課程のあり方の議論に資する大規模学力調査について考察する資料を提供することを目的とする。

研究の経過

1．平成 11 (1999) 年度に文部省「教育改革の推進のための総合的調査研究」において，研究課題「アメリカにおける学力調査の調査分析」の助成を受け，平成 12 (2000) 年 4 月に実施機関 ETS（Educational Testing Service）を訪れて NAEP に関する調査を行い，サンプリング技術や問題の作成・分析技術等に関する情報を収集した。

2．その後，本研究の研究代表者である荒井を代表として，上記研究の研究分担者を主たるメンバーとする「全米学力調査研究会」を組織し，平成 12 (2000)年の訪問調査時に収集した資料を中心に精査を行った。さらに，テーマごとの分担執筆方式によって，調査報告書をまとめた。

研究の成果

本研究においては，上記の報告書に 2001 年以降の NAEP の展開に関する補遺を補足し，完成稿を作成した。報告書の構成，担当，内容の要約は以下の通りである。

序　（荒井）

当時，騒がれはじめた学力低下問題をきっかけに研究を開始した。わが国では学力低下を指摘する側も応戦する側も適切なデータをもっていない。早晩，NAEP のような緻密な学力調査が必要になる筈と考えた。

第Ⅰ部　NAEP の概要

第１章　　NAEP の概要とその背景（荒井）

最初に米国の接続とシステムとしての学力評価について，SAT と ACTAP を取り上げて歴史的に論じた。SAT は東部のアイビー・リーグと言われる私学名門校を中心とする入学者選抜に用いられた知能検査型の進学適性テストが起源であり，基本的にエリートの選抜を目的としたものであった。それが，第 2 次世界大戦後の G. I. 法の下で急速に拡大した。ACTAP は 1950 年代の終わりに 4 教科のアチーブメント・テストとして登場し，大学大衆化時代に対応した試験として一気に広がった。しかし，このような「棲み分け」は，近年，到達度重視の考え方が強まるにしたがい，崩れつつある。

一方，戦後のわが国の大学入学者選抜においては，大学進学率が 4 割に達しようとしていた 1970 年代の終わりに大規模アチーブメント・テストである共通 1 次が導入された。それが序列化批判の下，1990 年代に入って大学入試センター試験に替わり，多様化と私大の本格参加によって結果的に高校教育の細分化，断片化をもたらした。さらに，推薦入学と AO 入試の導入，拡大により，入試はミクロレベルの最適化に焦点化され，システムとしての入試の大衆化の達成には程遠い状況となった。

次に，本報告書の主題である全米学力調査 (NAEP) について論じた。NAEP は公民権法の制定と軌を一にするように 1960 年代の終わりに開始された。その時々の教育問題に焦点を合わすメイン NAEP と長期的な学力変化を追跡するトレンド NAEP の 2 種類がある。前者は，内容はフレキシブル，後者は同じ問題を使う。NAEP のデータ公開には最初は極めて慎重であったが，90 年代には州別データやより規模の小さい単位での公開が可能になった。各州で行われている統一テストと連動して，教育スタンダードの設定に一役買っている。

わが国では，戦後に 3 度，全国学力調査が試みられた。第 1 期は 1948～54 年で，研究的な色彩の濃いものであった。第 2 期は 1956～66 年で，行政主導であり，「学テ闘争」を引き起こした。第 3 期は 1981 年以降で，新規教育課程の導入に合わせて実施されてきた。

わが国では以前の経験が継承されず，その点で米国とは格段に技術的な開きが生じてしまった。その点では米国の専門的経験の蓄積に学ぶべきことは多い。

第２章　　NAEP の問題設計（倉元）

メイン NAEP とトレンド NAEP の双方について，教科ごとに問題設計の考え方をまとめた。

メイン NAEP は調査時点で重要と考えられる事項を網羅するように，その都度作られる。本章では 2000 年に実施された読解，数学，理科の例を紹介した。

読解では，読解過程を「読者，文章，読書体験の内容を含む動的で複雑な相互作用」として捉えている。測定目的は，「文学的経験のための読解」，「情報のための読解」，「課題達成のための読解」の 3 つである。

数学では，以下の 5 つの内容をカバーしている。「数の概念，特性，操作」，「測定」，「幾

何と空間の感覚」,「データ解析，統計，確率」,「代数と関数」である。

理科は以下の 2 つの大きな次元に沿っている。「科学の分野：地球，物理，および，生命科学」と「科学を知り，行うこと：概念的理解，科学的調査，実際的推理」である。

トレンド NAEP では，一度設定した問題設計を変えることは出来ない。

1983-84 年の調査で読解は「読んだことを理解する」,「理解を広げる」,「読書体験をまとめる」,「読んだものを評価する」，1984 年の調査で作文は「様々な目的で書く」,「作文過程をまとめる」,「書き言葉の形式を整える」,「作文活動や完成した作文を評価する」，1985-86 年の調査で数学は「数学の基礎的な方法」,「離散数学」,「データのまとめと解釈」,「測定」,「幾何」,「関係，関数，代数表現」,「数と操作」，1985-86 年の調査で理科は「生命科学」,「物理」,「化学」,「地球宇宙科学」,「科学史」,「科学の性質」を測定目的とするとされている。

第Ⅱ部　NAEP のテストデザイン

第３章　　マトリックス標本抽出法と BIB デザイン（村木）

NAEP で用いられているサンプリングの方法として，被験者としての児童と，広範囲なテスト項目の双方を標本原理に基づいて選ぶマトリックス標本抽出法がある。

被験者抽出には層化抽出法が用いられている。NAEP では多段抽出法を用いて，その各段階のそれぞれに，層化抽出法を組み合わせていく。地域と都市規模による 8 個のユニバーから第一次抽出単位が抽出され，第二次抽出単位として学校が抽出される。非公立の学校や少数人種が多く在籍する学校は，母集団における実際の割合よりも多めに過剰抽出される。各学校それぞれから，最終の抽出単位である児童生徒が選ばれる。選ばれた学校，各児童生徒の NAEP 測定への参加を強要することはできない。最近のアメリカの状況を反映して， SD (障害を持つ児童) / LEP (英語力が限られている児童) の児童生徒をいかに含めていくのかという問題が起こっている。

テスト小冊子は BIB デザインの下に編集される。テスト小冊子はいくつかのブロックから成り立つ。種類や組み合わせは，科目ごとに異なる。異なるブロックには同じテスト項目は含まれない。ブロックの種類ごとに問題の質や難易度，制限時間が違っている。BIB デザインでは，共通のブロックで全体として連結することが項目パラメタ推定の重要な条件となる。

第４章　　古典的項目分析と項目応答理論モデルによる項目カリブレーション（村木）

古典的テスト理論，および，項目反応理論に基づく項目の尺度化の考え方について解説した。

NAEP のテスト項目のほとんどは多肢選択式である。

集団間の学力の差異を除去した後，特定のテスト項目にあらわれた難易度についての差異項目機能分析 (DIF 分析) を行っている。特定の集団に対して，学力以外の要因が影響して不公平なバイアスがあるテスト項目は，分析から除外される。

作業式の項目もある。高度な思考が必要な問題の作成にも力をいれている。

解答構築式項目は，簡便型と拡張型に分けられる。空欄補充は簡便型の，作文は拡張型の典型である。解答構築式項目の特徴は，評定が必要なことである。NAEP では評定者の人選と訓練に力をいれている。有資格者にさらに訓練を行っている。その結果，評価の信頼度や評価者間の評価の合致度はかなり高い。

NAEP のデータ分析は，専門のスタッフが担当している。多肢選択式項目は採点機械で高速度に処理される。解答構築式項目の評価は経費と時間がかかる。

NAEP では一般化部分採点モデルという IRT モデルを，テスト項目の分析に用いている。多肢選択式の項目には 3 パラメタロジスティックモデル，簡便型記述式項目の二値反応には 2 パラメタロジスティックモデルが使われている。作業型，拡張型記述式項目は段階評価である。

第５章　　調査結果のスケーリングと推算値（斉田智里・村木）

項目の尺度化における NAEP の実際について解説し，被験者個人の能力推定を伴わずに集団の学力分布を推定する推算値 (Plausible Value) の考え方について紹介した。

上述のように，NAEP では，項目の種類や採点方法に応じて 3 種類のスケーリング・モデルが用いられている。すなわち，3PL，2PL，GPM である。線形変換の範囲で原点と単位とが任意に定められるが，多くの場合，成績は，科目毎に受験者能力値 θ の尺度を，0 から 500，または 0 から 300 の尺度に線形変換をして報告される。

NAEP の目的は，全米の児童生徒の学力分布とその経年変化を明らかにすることであり，各受験者について調査するものではない。受験者とテスト項目の両方をサンプリングして学力分布を正確に推定する方法は推算値法と呼ばれている。

NAEP では θ を欠損データとみなし，t の近似式を求めている。

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

各受験者について，項目応答 x と背景変数 y が与えられたとき，学力の予測条件付分布から無作為抽出により，t^* の値を求めることができる。条件付期待値が推算値 (P.V.) である。t 値の計算過程で能力値 θ は，受験者の条件付分布から無作為に選ばれた値と置き換えられる。したがって，個々の受験者の能力値 θ を求めることは不要となる。

第６章　　Writing の評価（平井洋子）

1998 年の資料を基に，Writing の評価方法と調査結果に関して報告した。

メイン NAEP の課題には叙述型，情報提供型，説得型の 3 種類が設けられた。採点基準は ETS によって作成され，確認の後，予備調査で修正が加えられた。採点基準は，学年ごと，課題の種類ごとに作成されている。採点者間および採点者内の非一貫性をなくすため，次の 3 つの方策がとられる。監督官による抜き取りチェック，基準合わせ，採点者間信頼

性の計算である。尺度得点は生徒が「どの水準まで書く力が到達しているか」を表すものであるが，学年ごとに基本レベル，中級レベル，上級レベル (Advanced) という 3 つの到達度レベルが設定された。基本レベル未満の生徒が 16%から 20%いた。

トレンド NAEP は課題や回答例が公開されておらず，具体的な評価のされかたはメイン NAEP ほど明らかではない。

第７章　芸術科目の調査（Performance Assessment）（池田央）

1997 年の資料を基に，芸術科目の評価方法と調査結果に関して報告した。

芸術の枠組みは，創作，実技，観賞の 3 つである。演劇は創作と実技で 3 ブロック，美術は創作のみで 3 ブロック，音楽では，創作と実技の 3 ブロックのほか，何かの演奏活動に従事している者は 2 ブロック多く，観賞はどの科目も 4 ブロックの課題構成からなる。IRT 尺度は一般化部分採点モデルを用いて，平均 150，標準偏差 35 になるように尺度化されている。

第８章　NAEP を軸とする学力テスト・調査の技術革新：教育調査としての NAEP（池田）

NAEP の目的から調査の設計，調査方法，尺度化の方法，信頼性の確保，継続調査としての技術，背景情報，報告書の作成，携わる専門機関に至るまで，技術的な革新に関して解説した。

調査方針は NAGB (National Assessment Governing Board) で決められる。調査すべき教科が選ばれ，教科ごとに内容の枠組みが作られる。NAEP の目的は生徒個人の学力測定ではない。しかし，親や教師にとっての関心は，自分が関わる特定の子や学校についての情報である。それについての特定化は出来ない。質問したい項目が多岐にわたるのでマトリックス標本抽出法が開発された。結果を比較するため項目応答理論が利用されている。得点は回答者サンプルが抽出された過程を考慮して統計的に補正した形で算出し，その数値を比較に利用する。これを「推算値手法」と呼ぶ。NAEP で独自に開発された手法である。解答構築式問題の採点の信頼性を確保するために多大の努力を払っている。大規模調査は費用がかかるからといって，単発的に行ったのでは統計資料として不充分である。大規模継続調査制度を実施する米国の姿勢は敬服に値する。メイン NAEP とトレンド NAEP を用意することで，新しい変化と連続性という矛盾する調査目的の課題を解決している。州別 NAEP は全国尺度と共通尺度に等化され，比較できる。NAEP では背景となる比較関連情報を生徒，教師，校長から集めている。膨大な統計資料は利用したい対象者向けに整理されて，公開されている。調査の各ステップで，それぞれの専門機関が役割を発揮して全体が支障なく実施されるよう配慮されている。

第Ⅲ部　NAEP の調査結果について

第９章　Reading と Writing（平井）

1997 年の資料を基に，Reading と Writing に関するトレンド NAEP の評価方法と調査結

果に関して報告した。

Reading の 9 歳では 1970 年代に得点が上昇したがその後は変化がみられない。13 歳では，わずかずつではあるが得点が上昇する傾向にある。17 歳では 1970 年代から 1980 年代にかけて得点が上昇する傾向がみられたが，その後は頭打ちかやや減少する傾向にある。どの年齢においても，より頻繁に読書する生徒のほうがほとんど読まない生徒に比べて，一貫して得点が高かった。

Writing では，11 学年で全体的に尺度得点の低下傾向がみられたほかは，8 学年，4 学年のいずれも得点の長期的な変動傾向はみられなかった。今後の NAEP では，コンピュータによる回答入力も検討課題のひとつとなる。

第 10 章　Mathematics と Science（倉元）

1994，96 年の資料を基に，Mathematics と Science に関するトレンド NAEP の評価方法と調査結果に関して報告した。

Mathematics の 9 歳では 80 年代に急激に達成度を伸ばし，90 年からは安定した傾向である。13 歳では，1978 年から 1994 年まで徐々にスコアを伸ばしている。17 歳は，1973 年から 1996 年まで緩やかな上下がある。効果的な授業には最低限の知識の教授が必要であり，やみくもに生徒の自主的活動を促したところで実効は少ない。

Science の 9 歳では 1969 年から 1973 年に達成度を落とした後，1982 年までは横ばい，1986 年から 1992 年の間に回復した。13 歳では達成度の低下傾向が 1977 年まで続いたが，直後から徐々に持ち直し，1992 年で 1970 年の数値をやや上回るまでに回復した。17 歳では，1969 年から 1982 年に 22 点もの低下を示した。その後，年々回復してきているが，1969 年時点の水準には達していない。

人種間の学力差はかなり大きく，差の解消の方向に大きく動いていると言いがたい。

最終章　NAEP から何を学ぶか（池田）

NAEP に関する調査結果を基に，わが国における教育関連情報の収集と公開の必要性に関する提言を行った。

国の教育方針や政策の決定には客観的で定量的な統計指標を必要とする。NAEP では最新の測定技術と方法を投入することにより，複雑な手続きを経て NAEP の個別結果を相互にリンク統合し，意味のある解釈可能な尺度を構成するようにしている。NAEP では，教育情報として必要な調査が持つ 2 つの矛盾する目的をメイン NAEP とトレンド NAEP の別々の調査を計画することで対処している。調査のネックとなる実施と集計には，近年の発達した情報通信技術の導入によって解決の道を考えるべきである。情報が多くの学校端末から容易に集められ，集計できるシステムを開発することが有益である。個人情報の収集と扱いについては，きちんとしたルール作りが大切である。調査の実施機関は中立な第三者機関 NPO 組織が行うのが適切である。我が国が情報先進国に仲間入りするには，大学や研究所における研究者養成プログラムの充実も真剣に考えなければならない。

補遺　2001 年以降の全米学力調査（NAEP）の動向（木村拓也・倉元）

NAEP の歴史を 3 期に分け，主に 2001 年の NCLB 法 (No Child Left Behind of Act) 制定後に変質していった第 3 期の NAEP について解説した。

第 1 期 NAEP (60～70 年代) は，政策立案に資するデータの供出道具であった。しかし，目的は政策立案に資するデータの供出と定めながら，議論の末に決定されたデザインが妥協の産物であったが故に，非常に「鈍い」全国調査として誕生した。

第 2 期 NAEP (80～90 年代) は，自身に対する説明責任が問われた。1983 年の「危機に立つ国家」以降，連邦政府の教育予算に対する説明責任を果たす役割も付加され，調査デザインを「より敏感な」ものに変更していくこととなった。1990 年からの州別 NAEP の解禁が多額の費用を要する NAEP 自身に向けられた説明責任の応答であった。90 年代を通して，NAEP は機能拡充の議論にさらされることとなった。

第 3 期 NAEP (2001 年以降) は，州の説明責任を満たす道具としての役割を担っている。NAEP 自身の説明責任を問う議論が沈静化するのは，皮肉にも，NAEP 自身に州の説明責任を果たす道具としての機能が付与され，その性格において更なる転換を遂げた時であった。NCLB 法と NAEP が結び付けられたのが第 3 期 NAEP の最大の特徴である。

学力調査にどのような機能を付与するかということで，その効果を正にも負にもはじき出せることが，NAEP を巡るアメリカの経験から導き出される。わが国における学力調査も，目的，機能，その組み合わせの議論が重要になってくる。特に，鍵となるのは調査デザインであり，その設計は現在の技術水準の限界に応じたものとなる。コストパフォーマンスの問題も考えていかなければならない課題と言えよう。

教育ネットワーク研究室先端的プロジェクト型研究A型の補助を受けて編集,印刷をした本報告書を関係各方面に頒布したところ，基本的に好評であった。今後は，本研究の成果を基にして，わが国における大規模学力調査の方法について具体的な提言に結びつく研究を行っていきたい。

研究代表者　　荒井　克弘
（報告執筆責任　研究分担者　　倉元　直樹）