

I・HEAP (New Directions in Assessing Historical Thinking)

9章：歴史の知識と推論に関する大規模評価

全米学力調査の米国史評価

担当：池尻良平（東京大学大学院情報学環）

ikejiri@iii.u-tokyo.ac.jp

著者

・ Stephen Lazer

プリンストン大学で政治学の修士号を取得し、世界政策研究所の政策アナリスト、プリンストン大学の講師を担当。その後、1985年に Educational Testing Service に入社し、1991年から全米学力調査を担当。現在は、ETSの子会社の Questar Assessment の社長兼 CEO。American Educational Research Association (AERA) と the National Council for Measurement in Education (NCME) のメンバーも務めている。専門はアセスメントのデザイン、グループスコアのアセスメント、コンピュータベースのテストなど。



■用語

- ・ the National Assessment of Educational Progress (NAEP) | 全米学力調査
→実際の問題は「<https://nces.ed.gov/nationsreportcard/>」を参照
- ・ Matrix-sampled assessment | マトリックス・サンプル評価

■議題

- ①マトリックス・サンプル評価は日本でも普及可能か（歴史総合や調査形態と比較して）
- ②無解答の割合を減らす（意欲的に自由記述に取り組ませる）にはどうすべきか
- ③自由記述の作問のコツは何か
- ④デジタル化することで歴史家の史料を探す能力まで測定できる点は面白いが、現実的にどう設計し、どういう基準で評価できるのか

■概要（イントロダクションの第2段落より）

全米学力調査（the National Assessment of Educational Progress（通称 NAEP））の概要をまとめている章。はじめに、全米学力調査における知識とスキルのフレームワークについて説明している。次に、全米学力調査で歴史的思考を評価する際の機会と限界について議論している。続いて、全米学力調査での問題例を挙げている。最後に、全米学力調査における今後の課題について議論している。

■イントロダクション：全米学力調査(p.145)

- ・全米学力調査は、様々な学問領域の達成度を継続的に調査する唯一全国調査である。
- ・全米学力調査の主目的は、児童・生徒が何を知っていて何ができるかを測定し、達成度の傾向を追跡的に調査すること。
- ・全米学力調査はグループスコアの評価であり、個々の生徒の結果は報告しない。
- ・全米学力調査は、マトリックス・サンプリング・アプローチに基いており、児童生徒は長時間で総合的な評価の一部を受ける形式になっている。
- ・ほとんどの科目では、第4学年、第8学年、第12学年で実施される。
- ・本章では全米学力調査の概要を提供することを目的としている。

■全米学力調査の米国史のフレームワーク (pp.145-147)

- ・全米学力調査で測定する歴史の知識とスキルのフレームワークは1994年に作成された。
- ・フレームワークは過去20年間更新されてきたが、基本的な要素はそのまま残っている。
- ・「フレームワーク」の文書中には、テストの開発者に対して以下の4つのテーマをカバーする問題を作成するように指示している（時系列のカテゴリでも評価されるが、評価結果の推定が行われるのはテーマ別分野）。
 - ①アメリカの民主主義の変化と継続：思想、制度、出来事、主要人物、論争
 - ②人々、文化、思想の集まりとインタラクション
 - ③経済と技術の変化、その変化と社会、思想、環境との関係
 - ④世界におけるアメリカの役割の変化
- ・知識とスキル（探究、分析、推論）のバランスを取ることも問題作成者に課せられた内容。
- ・「フレームワーク」の作成者は、推論のスキルを明確に重視していたが、一方で知識の評価が不要になるとは考えていなかった。
- ・全米学力調査は、既存知識を測定する問題、分析と解釈のスキルに焦点を当てた問題、および推論と知識を一緒に測定する問題を含むように設計されていた。
- ・全米学力調査では、歴史の知識と歴史的思考の認知プロセスを以下のように定義。

歴史の知識と視点

歴史上の人々、出来事、概念、テーマ、運動、文脈、歴史資料、順序づけられた出来事を知り、理解すること。複数の視点を認識し、異なる集団の目を通して時代や運動を見ること、米国史の一般的な概念を発達させること。

歴史の分析と解釈

歴史上の問題を説明すること、歴史のパターンを特定すること、因果関係を確立すること、価値の言及を発見すること、重要性を確立すること、歴史の知識を応用すること、より良い結論を導くためにエビデンスの重みづけをすること、妥当な（defensible）一般化をすること、洞察的に過去を説明すること

■全米学力調査における歴史の知識と推論の評価：機会、限界、キーデザインの決定 (pp.147-152)

- ・大規模調査評価、特に全米学力調査では、評価設計者にユニークな機会を提供すると同時に、特定の課題をもたらししている。これらの機会と課題には、グループスコア評価の構成が関連している部分と、歴史の測定に伴う特有のものが関連している部分がある。
- ・生徒個々のレベルでスコアを得る歴史の全てのテストでは、テスト範囲と解答時間の問題を抱えている。
 - 様々なテーマ、地域、時代に関する問題を含める方が信頼性の高いテストになる
 - 多くの問題を出題する結果、各問題の解答時間が短くなる傾向になる
 - 推論スキルに焦点を当てた広範な刺激を与えるテストを提供したとしても文脈は重要
 - =ある時代の一次史料を読解できても他の時代の一次史料を解釈できるとは限らない
- ・解答に長時間かかるパフォーマンス課題を用いる評価を行うことで、他の方法では測定できない推論スキルが測定できるメリットがあるものの、これらの問題は出題範囲を圧迫し、スコアの信頼性に影響を与えるデメリットがある。
- ・全米学力調査のようなマトリックス・サンプル評価は、この問題に対して柔軟な対応ができる。
- ・第8学年の米国史の評価の場合、テスト問題全体は250分の分量があるが、テスト自体は25分を1ブロックとして10ブロックから構成されており、これらから2つのブロックを受験する（つまり、テスト自体は50分）。テストの冊子は全ての2つの組み合わせを考慮して50冊分作成される。
- ・全米学力調査の第8学年と第12学年向けの米国史評価では、資料を読んで1つのトピック（世界恐慌など）の質問に答える「テーマブロック」が含まれている。
 - 領域横断的な能力と焦点化された領域の能力のバランスを取ろうとしている。
- ・テーマ別ブロックをマトリックスに追加することで、集合させた評価は長時間のものになるものの、個人のテスト時間は長くないようになっている。
- ・これにより、全米学力調査の米国史評価は、他の一般的な歴史のテストに比べて、自由記述形式の問題の割合をより高くすることができる（表9.1参照）
- ・全米学力調査の米国史評価の全国サンプルは、第8学年と第12学年では1万人おり、約2000人が1つの項目を回答するため、より効率的に学力の推定ができる。
- ・全米学力調査では、児童・生徒、教師、学校の管理職にアンケートも実施しており、教授法、生徒の興味、学校の方針に関するデータを成績と関連付けることもできる。

表 9.1 2010 年の全米学力調査の米国史評価のタイプ別項目の配分

	多肢選択	短文の自由記述	拡張された自由記述	トータル
第4学年	63	28	4	95
第8学年	118	37	11	166
第12学年	110	36	13	159

- ・一方、全国的な歴史カリキュラムがないため、児童・生徒はテストを受けるための学習の準備ができず、点数が低いという問題点がある。
- ・そのため、DBQのような小論文問題や長文問題は出さず、テーマブロックでは関連する資料のセットを出しつつ、短めの自由記述形式の問題を出している。
→それでも肢選択に比べて無回答や省略される割合は高く、問題になっている。
- ・生徒の成績の変化を測定するために、全米学力調査では共通ブロックを使用して経年調査をしている。
- ・自由記述形式の問題の達成度の傾向を正確に報告するために、採点者は新しい採点が前の年の採点と異なっているかどうかを確認され、異なっている場合は再度採点を行わせるようにしている。この採点傾向のチェックは評価者の一致率以上に重視されている。
- ・アメリカの第4学年は州によって歴史で学んでいる内容がバラバラで、共通しているものが少ないにも関わらず、第4学年にも全国学力調査を実施している。
→何を知っていて、何ができるかを測定することを目的にしているため
→知識の応用を求める問題の場合、解答に用いる内容を選択できるようにするなどの工夫がされているが、課題は多い。

表 9.2 2010年の全米学力調査の米国史評価の無回答 (omitting) の平均パーセンテージ

	第4学年	第8学年	第12学年
多肢選択	1.59	0.75	0.67
短文の自由記述	6.76	7.85	9.43
拡張された自由記述	18.70	5.86	6.74

■全米学力調査の米国史評価の問題例 (pp.152-155)

◎第4学年 (pp.152)

- ①簡単に地図を解釈させ、既有知識を問題の解答に応用させる問題 (図 9.1)
 - 問題1は1810年の地図から2つの仕事を書き出す問題(28%が満点、56%が部分得点)
 - 問題2はなぜ多くのまちが川の近くに位置したのか、その原因を2つ書く問題
(15%が2つ挙げられ、59%が1つのみ=やや難しかった)
 - 「経済と技術の変化」のテーマ(の知識)と、「歴史の分析と解釈」の認知領域を測定
- ②19世紀の日記をもとに、今の子どもたちと何が違うかを解釈させる問題 (図 9.2)
 - 35%が満点、14%が部分得点

◎第8学年 (pp.152-156)

- ①移民の割合を描いた2つの円グラフを読ませ、円グラフの正しい解釈をさせる問題と、2つの円グラフの変化の歴史的な理由を述べさせる問題 (図 9.3)
 - 前者の問題は正確に解釈できた

→後者の問題も解けた満点の生徒は15%のみ(=違いを説明することは難しい)

- ②フィラデルフィア憲法制定会議での代表権のあり方をめぐる論争について説明し、この論争を解決した「Great Compromise」について説明させる古くからある問題
→最高点はわずか6%で、24%が部分点。20%が無回答(=かなり難しい)

◎第12学年 (pp.155-156)

- ①政治的な漫画を解釈させ、歴史的背景を説明させる問題(図9.4)
→「世界におけるアメリカの役割の変化」のテーマ(生徒は苦手なテーマ)
→63%が0点、19%が無回答/課題外の回答(=非常に難しい問題)
- ②農業規模が増加している簡単な量的データを解釈させ、変化の原因となった発明を書かせる問題(刺激材料の解釈と既有知識の応用の両方が求められる問題)
→29%の生徒は表を正確に読み、変化のパターンの歴史的説明もできたが、45%は前者しか解答できていなかった(中程度の難易度といえる)

■将来の全米学力調査 (pp.156-157)

- ・1992年と1993年に開発され、1994年に実施された全米学力調査の米国史評価は、基本的な構成は変わっておらず、多肢選択問題と自由記述の問題の組み合わせで構成された紙のテストである
- ・しかし、2014年の調査終了後、ほぼデジタルに移行し、内容も大幅に変更している。
→フィルムやビデオの動的な刺激も可能に
→将来的には多様な資料(テキスト、画像、ビデオ、音声など)を用いた質問も可能で、歴史の事実を知るだけでなく、どこでどのようにその事実を見つけられるかを知ることとも可能になると思われる
- ・シミュレーションやゲーム、没入型の環境、コラボレーションの評価も考えられる。
- ・人工知能による自動採点の可能性もあり、コスト削減、傾向性の測定の容易化につながる。
- ・テクノロジーを使ったテストによって、知識の測定を重視するという前提は疑問視されるようになるかもしれない。
→必要な情報を素早く効率的に見つけ出し、良い情報源を見分ける能力が測定対象になるかもしれない。
- ・いずれにせよ、グループスコアの評価のメリットは意識すべきで、継続すべき。

■まとめと結論 (pp.157-158)

(上述した内容を言い直しているだけなので省略)