

17 章：歴史教育における学習評価：過去,現在,可能性のある未来

Assessment of Learning in History Education : Past, Present, and Possible Futures

担当：玉井 慎也（広島大学大学院）

山村 向志（千葉県立姉崎高等学校）

I. 著者情報（写真や経歴などは WEB 上で発見できず）

著者名：Denis Shemi It

研究関心：教育評価,歴史教授と歴史学習を構成主義的アプローチで捉えること

経歴：the University of Leeds に20年以上勤務。1986年からTrinity and All Saints, a constituent college of the University の the School of Education における学部長を務めている。1974年に13-16歳を対象にした the Schools History Project の評価を担当し,1985年には the Cambridge A' Level History Project の共同ディレクターを担当。近年では,教育のマネジメントやビッグ・ヒストリーの教授、学習に関心がある。

（“knowing teaching and learning history” 著者紹介欄より）

II. キーワード

- ・ selected-response items 選択式の解答問題
- ・ constructed-response items 記述式の解答問題
- ・ declarative knowledge 宣言的知識

III. 議論したいテーマ

- 方法的概念は、それらの一つずつ（観点別に）評価することができるのか（それとも、複数の方法的概念を巻き込んで評価ルーブリックを作成して見とることが妥当なのか）
- ペーパーテスト（リフレクションシート）で歴史的思考・歴史的意識をどのように評価していくことが可能か

IV. 本文の要約

<はじめに>：歴史教育における評価の現状と課題（pp.449-450）

- 20世紀半ばには,大西洋の両側にある学校の歴史授業のほとんどは,国や地域の過去の知識が伝えようとしていた（Wi lschut,2012）
- 従来のアプローチは,多くの教師が目標をほとんど顧みずに「評価するべきものではなく測定できるもの」を評価するものだった。
- こうした批判は,新たに確立された評価方法の欠陥によるものではなく,実験データの蓄積や歴史教育の目的に関する新たなアイデアに遅れをとっていることによるものである。
- 評価方法の決定の際には「評価する対象」と、「評価の目標」を考慮に入れる必要がある
- また,評価のシステムの有用性は,4つの要因間の相互作用によって左右される。
 - ⇒（1）評価のデータが提供する目的（2）評価の対象となる学習の構成要素（3）データを測定する方法（4）データの信頼性やリソースの量の度合い。
- ・ 以下ではこれら4つの要因とその相互作用について説明し,評価の有用性とそれらの整合性について説明する。

<評価の目的> (pp.450-451)

- 生徒の学習成果に関する情報は、さまざまな目的別で求められる
⇒ 主に「診断的評価」、「形成的評価」、「総括的評価」の三つ
- 「総括的評価」には、学習コースの終了時の個々の学生の学習成果に関する情報が必要とされる。
- 「形成的評価」は、現在進行形の学習に関するエビデンスを用いて、進行中の教育の目的、方法、およびそのペースの把握を行うことで達成される。
- 総括的評価は学習の成果を評価し、形成的評価は教育目標がどの程度達成されているのかを測るが、「診断的評価」は学習における障壁や方向性の両方を明らかにする。

< 学習の構成要素 > (pp.451-453)

- 歴史を教えることの目的は、いまだに「受動的な」過去や選択された情報の伝達であり続けている。
- しかし、生徒が自身の知っていることを理解し応用することができない限り歴史教育の価値は疑わしいという理由から、歴史的知識の教授と評価は再考されてきた。(Lee,1991)。
- このような懸念は、学生が過去についての情報を理解し、評価し、応用するために必要な概念装置を定義する努力を促した。
⇒ 歴史的知識 (HK) ,歴史的思考 (HT) ,および歴史的意識 (HC)
- 歴史的知識には歴史に関する宣言的知識の認識(= 歴史的な説明や主張)や想起(= 歴史用語や歴史的な事実)が含まれる。前者は記述式の解答、後者は選択式の解答によって評価される。
- 歴史的思考は証拠の情報源を参照して、過去に関するより正しい解釈を峻別しようとする。また「なに」、「どのように」、「なぜ」といった、歴史的な問いを立てることも必要とされる。歴史的思考に現在焦点が当たることは、評価のための新しい方法が必要とされていることを示している
- 歴史意識は、歴史的知識と歴史的思考がわれわれの日常における信念、態度、意思決定に影響を与える方法を決定している。歴史意識に関して、生徒に何をどのように教えるべきかについての示唆が引き出されてはいるが (Lavi lle,2004) ,この構成概念の構造の規定にはまだわずかな進歩しかなされていない。そのため、歴史的知識と歴史的思考のより複雑な側面を有効に評価することが出来ない限り、歴史意識の信頼性のある評価は不可能であるかもしれない。

<評価のデザイン> (pp.451-453)

- 学習環境に即して行われる「学習課題(course work)」は定期テストよりも真正で、生態学的にも有効な評価の形式の一つであるが、それらのような「統制下におかれていない」環境で行われる評価の欠点としてデータの信頼の低さがある。
⇒生徒の学習成果は性別や社会経済的状況といった環境に影響を受ける
(そのためイギリスでは筆記試験と学習課題の二つを試験のスコアとして設定)

- ▶ また、プロジェクトや学習課題といったアプローチは、課題の比較可能性と成績の信頼性の優先度が低い場合（評価データが形成目的で使用される場合など）に役立つが、個々の生徒に対する成績づけがハイスティクスな決定となる際には不十分である。したがって、一般的に選択式解答項目（SRI）と記述式解答項目（CRI）が学校で用いられることは驚くことではない。

<選択式の解答問題> (pp.454-457)

- ▶ 選択式の解答問題(以下SRI)には、「正誤選択」、「穴埋め」、「組み合わせ」、「複数選択」を含むさまざまな形式がある。すべての生徒は、2 つ以上の単語やステートメント、またはリンクのうち、どれが与えられた問いに対する「正しい」答えであるかを特定する必要がある。
- ▶ テストの解答範囲が限られており、正確性が優先されるため、成績付けは人間よりも機械によって確実に行われる。
- ▶ 問題なのは、歴史的思考と歴史的意識のSRI ベースの評価であり、生徒の決定手順の分析（Reich,2015）から、歴史的思考に関わる推論が多肢選択項目への応答からは引き出すことができないことが示唆されている。
 - ⇒ 複雑な説明や主要な情報源に関する選択肢を選ぶ場合、正しい間違っ了解答がなぜ行われたのかという理由や、学習したこと/学習していないものに関する情報を得ることができない。
- ▶ 歴史的思考の評価に適したSRI を開発する試みも行われている。1 つの提案は、「重み付けされた多肢選択項目（VanSledright,2015）である。
- ▶ 以下の項目は目撃者の証言の限界に焦点を当てている。ここでの重み付けは、目撃者の証言が物的な証拠と比較して信頼性の低い証拠であることを示している。全員が死亡したことはわかっているが、リトル・ビッグホーンの戦いでカスターと彼の軍隊に何が起きたのかを理解することには困難が伴う。なぜなら..

- a) 生き残ったインディアンは、証言を提供するほど十分に英語を話せなかったから。（0）
- b) カスター大佐と彼の軍隊が殺され、彼らが戦いについて語るができなくなったから。（1）
- c) ベンティーンとリノは戦いの様子を聞いたが、起伏のある地形だったため見るができなかったから。（2）
- d) 証拠の消失により、戦場の様子を再構成することが阻まれているから。（4）

- ⇒ しかし、重み付け(0)の選択肢＝ 選択肢(a)は事実上「誤った」選択肢であるが、インディアンで英語をしゃべることができた「スー」が生きていれば、選択肢(b)のカスターの軍隊からの証言の信頼性に匹敵する可能性があることを認めない。
- ⇒ 項目の重み付けの有効性は、テストを行う際に生徒に「口頭で報告させたり」や「教室で議論させたり」することにより確認できる。

<記述式の解答問題> (pp. 457-460)

- 記述式の解答問題(以下、CRI)は特定の質問に対する1語または2語の回答を提供するか、エッセイ形式で教えられたコンテンツについて知っていることをすべて書くことから始まる伝統がある。ほとんどの国では、宣言的知識(declarative knowledge)を評価するCRIが知識の適用と想起を要求する評価やエッセイの長さの評価に取って代わり、歴史的思考や歴史的意識を活性化しようとする評価に移行している。後者の評価でのCRIは子どもを刺激する資料を中心に記述され、歴史的なパラドックスやジレンマ、予期せぬ出来事の変化、奇妙な慣行についてのコメントを求める(例: 英国の16歳以上のCRI)。

- ① 患者の尿を調べる10世紀のヒポクラテスの木版画の写真(刺激資料)を提示する
 - ② 写真に居るヒポクラテスがアラブの服を着ている理由を説明するように求める
- 刺激資料は、ヒポクラテスが紀元前5世紀のギリシャの医師であったという想起された知識を活性化することを目的としている。刺激資料は、子どもの既存の木版画のイメージと矛盾する。これによって歴史的思考の変化と進歩に関連する証拠を見とる。

【2つの回答】

回答1: ヒポクラテスはアラビアを訪れ、アラブの服を着ていたに違いない。

彼の写真は10世紀の歴史教科書の記述にある。

回答2: この資料は、アラブ人がギリシャ医学から学んだことと尿を使って感染症を診断するといった古い考え方がうまくいったことを証明しているため、1000年以上に渡って重要な存在であり、今日でも機能している可能性がある。何百年の間、アラブ人はヒポクラテスがどこから来たのか忘れていたかもしれないし、ギリシャではもはやピポクラテスなんて(歴史的人物として)呼ばれていなかったかもしれない。もし彼がアラブ医学にとって重要ならば、彼はアラブ人であったに違いないと考えたかもしれない。

- 回答1の生徒は、ヒポクラテスがギリシャの服ではなくアラブの服を着ることを選んだ理由を説明するときに、紀元前5世紀という知識を以前の選択肢として概念化する。これにより時間の問題を棄却する。
- 回答2の生徒は、医学(および物質的・象徴的文化の他の側面)が時を経て発展したと仮定し、医療行為の連続性が知識の不連続性を伴っていることを示唆する。これにより想起された知識と提供された刺激資料との矛盾を説明する。
- このようなCRIは、生徒の歴史的思考に関する有用な証拠をもたらすが、完全に有効になるわけではない。一部の生徒は「知らない」とか「教えてないから答えられない」としてCRIに答えようとすらしめない。また他の生徒は主題や文脈について知っていることを書き最善を祈っている。より深刻な生徒は、何が問題なのかに対する誤解をしている生徒である。例えば、ある生徒はヒポクラテスがギリシャ人にはなれなかったと自信を持って断言した。「木版画から明らかに、ヒポクラテスはアラブ人であったことがわかるからだ!」

- ▶ 回答2のような生徒はHT「証拠」の観点から解釈し、ヒポクラテスの死後数世紀を経た信頼性の低い二次資料として木版画を却下したが、何が問題なのかに対して誤解をした生徒はHT構造の異なる次元、つまり、現実と代表されるかこのあいだの存在論的な関係についての考えと仮定を扱う「説明（account）」の次元についての歴史的思考の証拠を与えた。
- ▶ 宣言的知識の評価をすることを目的とするCRIは穴埋め式SRIの形式に似ているが、決定的な違いはCRIの評価の場合、専門審査官の裁量で評価されることにある。（例：「アメリカは__ __ __ にちなんだ命名された」における「Amerigo Vespucci」の記述で、「Vespucci」や「A. Vespucci」は点数を与えられるが、「Amerigo Vesconti」は容認できない。→選択肢を選ぶだけのSRIと精密な記述が要求されるCRIの違い）
- ▶ 複雑な歴史的知識の評価と適用を求めるCRIの場合、マークスキームでは複数の基準に対して生徒の回答を採点する必要がある。この基準は、ブルームタキソノミーの4つの段階から派生している（Bloom, 1976）。（例：「1798年から1799年にかけての出来事は、フランス国民に自由、平等、友愛をもたらすことができなかった。なぜか。」）

- 1) 情報の選択：(a) トピック (b) 質問 (c) アーギュメント
- 2) 提供された文章を読解するための情報の使用：(a) 現代、自由、平等、友愛など18世紀後半の概念を区別できる信頼性のある情報 (b) 18世紀後半の概念に付随する様々な意味の例示 (c) 自由、平等、友愛の概念の変化の例（1798年から1799年）
- 3) 提供された文章を評価するための情報の使用：(a) 自由、平等、友愛の差別的な評価（differential evaluations）(b) 提供された文章への反対意見の評価 (c) 例外が提供された文章を無効にするポイントの検討
- 4) 情報を使用して因果的議論を展開：(a) 因果的議論を構築し実証するため情報を効果的に使用したことを証明 (b) 専門的歴史家によって進められた因果関係の議論を評価 (c) フランス革命の様々な段階に特有の原因とプロセスを特定

- ▶ 歴史的思考を構成する「証拠」「歴史的視点」「倫理的次元」を評価するために考案されたもの（Seixas, Gibson & Ercikan, 2015）は、学問分野としての歴史の分析からコア基準を導き出した。この基準は汎用性が高く、さらに歴史的知識、歴史的思考、歴史的意識を評価するための基準を考案することに利用できる（Eliasson, Alven, Yngveus, Rosenlund, 2015）が、採点基準はア prioriに決定されるため生徒の反応の根底にある歴史的思考や歴史的意識を表すことはできない（評価できるわけではない）代償を伴っている。
- ▶ 英国では歴史的思考の5つの側面（「説明」「証拠」「変化と進歩」「因果関係」「共感的説明」）に特有の学習プロセスを評価する枠組みを設計（Blow, 2011；Lee & Ashby, 2000；Lee & Shemilt, 2003, 2004, 2009, 2011）。（例：「変化と進歩」）

- 時間の経過に伴う変化と進歩（Change and Development Through Time）
- レベル1：出来事と同等と解釈される変化。継続性＝「何も起こらない」
- レベル2：2つの時点の差として解釈される変化。継続性＝「すべてが同じまま」
- レベル3：その時点で生活している人々にとって重要であり、重要な違いであると解釈される変更。継続性＝「いくつかのことは同じまま」

レベル4：その時点で「重要な変化」と「時間の経過に伴う変化」，つまり「即時的な歴史的意義をもつ変化」と「長期的な歴史的意義をもつ変化」を区別する。継続性＝「局所的（locally）／歴史的（historically）に意義をもつ」

レベル5：進歩は時間の経過に伴う変化と継続性のパターンとして解釈される。継続性＝

① 変化の率（rate）と方向が継続的な変化「トレンド」，② 変化の速度および方向の変化を伴う進歩「転換点」，③ 進歩の水準の中断「不連続性」。

- マークスキームを作成するアプローチは、「構成（construct）ベース」（Lee, 2006）と「規準（criteria）ベース」＝生徒の物語，議論，判断を専門的基準との対応で評価の2つがある。

<知識要素> (p. 461)

- 生徒が歴史的思考を行使したり歴史的意識を表示したりする能力は，試験項目で参照される歴史的事象について何を知っていて，何を知らないかによって制約される。英国ではコース終了時の評価により知識ドメイン全体でSRIの分析をランダム化するか候補者にCRIの選択肢を提供することで上記の問題を回避した。
- 生徒が所有するあらゆる知識を展開するように誘う以下の英国のSレベル試験では，社会的・政治的・経済的な危機の知識と因果分析が必要だが，特にそうではない。例：「小さな火花で大火は起こるが，それが原因ではない」（プラトン）という意見について，歴史の知識を使い賛成か反対かを議論しなさい。
- データベースを提供して評価するテストは，英国の実験的な公的試験（Shemilt, 1979）やカナダ，スウェーデン，ドイツで最近提案された（Ercikan & Seixas）。

<測定要素> (pp. 461-467)

- スコアリングは，学習の道筋を明らかにしたり理解の障害を特定したりする場合，個々のスコアや成績のリストよりもグループの達成度を説明するプロフィールの方が便利である。
- 歴史的思考の「継続性と変化」の概念に対して次のようなルーブリックを提供している（Ercikan & Seixas, 2011, p. 256）。これは形成的評価にも使用できる。

レベル1：変化を特定できない

レベル2：継続性と変化の両方について基本的な観察を行っている

レベル3：特定の履歴に基づく継続性と変化に関する正確な主張をサポートしている

レベル4：3 までに加えて洞察力に富んだ議論を提示し様々な視点から継続性と変化に関連する力の関係を認識している

- 英国SHPカリキュラム開発プロジェクトの議論では「医学と時間」を巡って使用される共感的理解（あるいは歴史的視点）の概念を評価する問題が作成された。

「ギリシャ人は個人的な清潔さの重要性を強調し，ローマ人は新鮮な水と下水道の建設に多くの金を費やした。これは，彼らが細菌によって引き起こされる病気を知っていたことを証明している。」上記の結論に賛成または反対する理由を説明しろ。

【2つの回答（両者賛成の立場）】

回答1：アスクレピオスがいるから。彼は人々に洗濯して運動するように言っていた。

ローマ人はクロアカ・マキシマという下水道を作り鉛管に水を配管した。時に鉛中毒になることがあった。今日の車の排気ガスや工場下水を川に流すことへの対策も同様。

回答2：ローマ人は非常に賢く、道路や高架道路を建設した。また、英国を征服しハドリアヌスの壁を築いた。その時、泥だらけの人々が病気になったことに気づいて、土の中に菌があったことを突き止めたのだろう。ローマ人が細菌を気にしなくなった後は、神が細菌を綺麗にしたり汚いものから救ったりしたと考えた。

- ▶ 回答1 はレベル1 として採点され（理由：過去と現在の類似性について実証できていないから）、回答2 はレベル2 として採点された（理由：その時代の常識的見方は投影されているから）。しかし、一部の教師や評価専門家はこのランキングに異議を唱えた（回答2 は問いの半分にしか回答しておらず、回答1 は関連する証拠を提供している）。このように、本章で取り上げたルーブリックのレベルは私たち教師がなぜ・何を教え評価したいかを表してはいるが、その尺度がどのような意味を持つのか、レベル1 よりもレベル2 の方がなぜ・どの程度点数が高いのかについて確信を持っていない。
- ▶ 歴史的思考の概念を巡る評価の現状は、一つの概念で評価ルーブリックが作成されているが、将来的には例えば証拠の使用に関する生徒の歴史的思考の評価を歴史的記述の性質・状態に関する評価に巻き込んでいくことが期待される。
- ▶ また、歴史的意識の性質を突き止めることは依然として困難だとされているが、北米とヨーロッパで研究が続けられている（Duquette, 2015）。

<未来の可能性> (pp. 467-468)

- ▶ ますますグローバル化する世界の中で「大きな」歴史へのシフトが予想されている。5～6の「共通コアHT」の有効な尺度を開発できれば、歴史的思考の評価が当たり前になる可能性がある。
- ▶ 構成主義的研究（Leoni & Ntamakil iro）と認知負荷理論（Paas, Tuovinen, Tabbers, Van Gerven, 2003）の進歩が続けば、知識適用課題の複雑さを制御することと、学習で何ができるか・学習したことを反映する方法で生徒の反応を評価することの両方が可能になるかもしれない。
- ▶ 今後の評価の動向は、教育政策上の優先順位や学術的な研究の成果、教員養成課の改革によって左右される。
- ▶ 賭けるに値する唯一の予測は、評価データの現在の多くの使用が将来にも続くということである。テストの成績は、生徒に報酬、インセンティブ、雇用や高校の選択肢を与え、確立された通過儀礼に貢献する。