キャンベルの法則について ： 測定評価を巡る構造的陥穽に関する考察

木村, 拓也
九州大学大学院人間環境学研究院 ： 教授

https://doi.org/10.15017/4773103

出版情報：九州大学教育社会学研究集録. 23, pp.57-64, 2022-03-15. Seminar of Educational Planning, Measurement, Evaluation, Department of Education, Graduate School of Human-Environment Studies, Kyushu University
バージョン：
権利関係：

キャンベルの法則について

－測定評価を巡る構造的陥穽に関する考察－

About Campbell's Law: A Study on the Structural Trap in Measurement and Evaluation

木村　拓也

1. 問題の所在：測定評価を巡る問題構図

評価には，教育場面で想起されやすい人物評価だけでなく，機関評価や政策評価も存在する。また，多角的な評価資料を収集するアセスメント（Assessment）を通して評価が行われる場合もある。視認可能な状況の把握に測定（Measurement）が必要とされない場合があるとしても，社会や組織など，眼前の状況を視認するだけでは到底把握不可能な範囲までの評価には，情報要約の手段としての測定が必要とされる場合がある。測定によって得られた数値単独では意味を持ち得ないが，価値づけされることで評価がなされる。この意味で，評価は，価値の問題につながる。また，そうした評価が差別を引き起こす場合，選抜や昇進の問題が絡めば，社会リソースの公平な分配についての社会哲学の問題でもある。さらに，情報集約に技術が用いられる場合には，その技術に対する専門家の判断と誰が見ても了解可能な数値の「客観性」とに齟齬が生じる可能性もあり，その場合，科学技術社会論の問題にもなる。

グーデンベルグの活版印刷の発明の影響について，マクルーハンは口頭文化と活字文化の違いについて論じたが，そうした 15 世紀の印刷技術を可能にした精神的条件について論じたのがクロスビーである。クロスビー（1997=2003）は，活版印刷が，機械時計・暦・地図・海図・天文学・楽譜・絵画の遠近法・簿記によって既に用いられていた視覚化の威信を高め，数量化という行為を加速させたものと論じた（Crosby，1997=2003）。さらに進んで，ミュラー（2018=2020）は，数量化され視覚化されたものを受け入れる精神的素地の存在は，時間的・空間的に了解不可能なことを認知可能にする手段であったはずの数量化を，それ自体目的にしてしまうことで，計測が単純な／可能なもののみの数量化，情報の改竄，基準変更による数値達成，を誘引する危険性も生じさせると論じた（Muller，2018=2020）。また，専門家の判断に不信が募ったときには，こうした我々の精神的素地が利用され，「科学的な」手順を遵守して測定されることで，数値が途端に「客観性」という社会からの信頼性を獲得する，といったことも起こりうる（Poter，1995=2019）。

このように評価そのものが社会的文脈の中で行われ，更にその評価結果が社会で活用されるという意味において，評価そのものが，我々の社会的信念を強化してしまう。グールド（1981=1989）は，知能を量として測定することによって，個人やグループの社会的価値を表すことができるとした生物学的決定論を包含する，頭蓋測定学や知能テストを歴史的展望のもとに考察した。そうした頭蓋測定学や知能テストが孕む論理は，単に生物学的決定論の論理に留まらず，それが社会的信念に影響を与え，その構築された社会的信念が，何を測定するのか，何の因果関係を採択するのか，その解釈を妥当だと感じるのかにまで影響を与えるので，生物学的決定論の論理を再補強する（Gould，1981=1989）。この意味で，そうした測定を基にした評価において，統計的なバイアスのみならず，社会的なバイアスをも入り込む余地があることは，数値を評価する際に価値が入り込む以上，原理上不可避であるとも言える。

前述したミュラー（2018=2020）の関心は，クロスビーが論じた数量化され視覚化されたものを受け入れる精神的素地の上で，現在，「測定執着（metric fixation）」（Muller，2018: 18 =2020，p.19）とミュラーが呼ぶ現象に及んでい

る。「測定執着」とは，その主な要素として，「個人的経験と才能に基づいて行われる判断を，標準化されたデータ（測定基準）に基づく相対的実績という数値に置き換えるのが可能であり，望ましいという信念，そのような測定基準を公開する（透明化する）ことで，組織が実際にその目的を達成していると保証できる（説明責任を果たしている）のだという信念，それらの組織に属する人々への最善の動機づけは，測定実績に報酬や懲罰を紐づけることであり，報酬は金銭（能力給）または評判（ランキング）であるという信念」が存在し，「それが実践されたときに意図せぬ好ましくない結果が生じているにも関わらず」，それらの信念が「持続している状態」であると問題視した（Muller, 2018: 18 =2020，p.19）。

数量化され視覚化されたものを受け入れる精神的素地の存在とそれらを客観視するがあまり，執着し過ぎることで引き起こされる矛盾は，上述の測定評価が，純粋な測定技術のみによらず，社会の中で構造的に生じるものであると捉えることができるのであれば，そのことに関して，社会科学における碩学の知に触れることも必要であることは言うまでもない。

2.「キャンベルの法則」の論理構成

ミュラー（2018=2020）の指摘それ自体は，鋭い洞察として評価できる一方で，こうした測定評価とそれがもたらす社会への影響については，幾度となく社会科学者によって取り上げられてきた古典的な話題でもある [1]。例えば，社会心理学者で 1975 年にアメリカ心理学会（America Psychological Association：APA）の会長も務めたこともある，当時ノースウェスタン大学に所属していたドナルド・T・キャンベル（Donald T. Campbell）が 1979 年に発表した論文で提唱したのが，「定量的な社会指標が社会的意思決定に使われば使われるほど，腐敗の圧力に晒され，監視すべき社会的プロセスを歪めたり腐敗させたりする傾向が強くなる（*The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social process it is intended to monitor.*）」（Campbell，1979，p.85）という，「キャンベルの法則」（Campbell's Law）である。ミラー（2018=2020）が，大学，学校，医療，警察，軍，ビジネスと金融，慈善事業と対外援助の事例を通じて「測定基準への執着」を論じたのは，この「キャンベルの法則」に対する事例を挙げた長い注釈であったし，ミュラー自身も著書の中で，「キャンベルの法則」を引用している（Muller, 2018: 19 =2020, p.20）。そして，「キャンベルの法則」は，いわゆる社会学でいうところの「意図せざる結果」の一種でもある。

予め断りをつけておくが，キャンベル（1979）の論文自体は，著名な心理学者として，計量的な分析手法に基づく政策評価の問題点を統計的に緻密に批評した論文であり，決して，計量研究それ自体の批判を主題とする論文ではない。「キャンベルの法則」として，測定評価がはらむ政治的な問題を論じたのは，時系列デザインや無作為化実験について詳しくその問題点などを 23 ページ中，17 ページ論じた後の 3 ページ程（残る 3 ページは注釈）につけられた最後のお小言に近い位置付けにあるものである（それが 40 年ほど経った現在にまで残り伝わっているのは皮肉としか思えない）。キャンベル（1979）は，計量研究による政策評価の技術を，比較可能な複数の環境で，且つ，その評価を実施する規模に即した資金のもとで，純粋に代替可能となる施策に対してのみ禁欲的に適用すべきものであり，決して，個人やグループを評価したりすべきでないと，計量研究による政策評価の使途を限定する必要性を説くことを主眼とした論文である。

そして，この問題意識は，キャンベル（1979）の論文の中でも言及されているように，ジョンソン大統領の「偉大なる社会」プログラム（1964-68）に関連した貧困などの社会問題に対する大規模な効果実証研究の存在から生じたものであり，教育学の文脈で言えば，コールマン報告（1966）やその再分析をしたジェンクスの『不平等―学業成績を左右するものは何か』（*Inequality: a reassessment of the effect of family and schooling in America*）（1972）の刊行された時期であるといえば，当時のアメリカ社会における問題意識として理解されやすいのかもしれない。

キャンベルは，社会実験として大規模に行われる実証研究がもたらす実証結果の真実性について，その社会実験であるが故に生じるバイアスを，心理学者らしく「外的妥当性」（external validity）の問題であると表現したり（Campbell，1979，p.81），実験を受けているということで特別意識を持ってしまったり，社会的に望ましい行動

をとりたがる「ホーソン効果」や「モルモット効果」，また，経済学者に多く見られた批判として，時間が制限されていることで参加者の行動をその時だけよく振る舞う傾向をもたらすといった「時間制限効果」があることも紹介している（Campbell，1979，p.81）。

そこからさらに踏み込んでキャンベル（1979）は，社会で生じる測定評価が社会にもたらす影響を，端的に，「定量的指標の腐敗効果（corrupting effect of quantitative indicators）」（Campbell，1979，p.84）と呼ぶ。そこで挙げられたのが「キャンベルの法則」の具体例である。まず，キャンベル（1979）があげたのは，アメリカ社会における投票統計と国勢調査の例であり，国勢調査は政治的意思決定に使われていないが，投票統計は国民の仕事や生活，権力に深く関わるが故に，不正が生じやすい状況にあるという（Campbell，1979，p.85）。二大政党政治であるアメリカ社会の中で，選挙によってどちらの党が勝利し，政権を握るかによって社会政策が左右されるアメリカならではの状況が背景にあることは明らかであろう。二つ目の例としてあげているのが警察の犯罪解決率の例である。犯罪解決率が評価の対象となると，現行犯逮捕者に未解決事件について，実際には犯行を行なっていないにも関わらずに，自白を誘導する圧力が生じることを紹介している[2)]（Campbell，1979，p.85）。これも司法取引があるアメリカの社会的背景を色濃く表すものであると考えられるが，だからこそ，この2つの具体例を読者に訴えかけるために取り上げたのであろう。

その他にも，「政府機関の労働者に設定された生産性基準が，プログラムの有効性に悪影響を及ぼす形で労働者の努力を歪めている」（Campbell，1979，p.85）として，事務スタッフを処理したケースの数で評価するようになると，迅速で効果のない面接や斡旋が横行し，処理するのが簡単な案件に努力が集中するおかげで，サービスを最も必要とする人を無視することになる，という矛盾も紹介している（Campbell，1979，p.85）。

また，工場生産を例にとり，製品の金銭的価値，製品の総重量，あるいは，生産する品目数が数値目標とされた場合，「工場の生産性を評価する公式な目標として用いると，生産性を機能的に歪めてしまう」（Campbell1979:86）として，ひとたび金銭的価値を数値目標に定めれば，1つの製品を作るためだけに工具を使い，労働効率を悪くして，わざと金銭的価値を釣り上げることに努力が傾けられ，ひとたび製品の総重量を数値目標に定めれば，最も重い製品だけを生産することに努力が傾けられ，ひとたび生産する品目数を目標数値に定めれば，最も制作労力のかからず，数だけ稼げる製品だけを生産することに努力が傾けられ，結果，皮肉として，「不要なものは過剰に生産され，必要なものは不足してしまうのである」（Campbell, 1979, p.86）と述べている。こうした事態は，単に1970年代のアメリカのみならず，エビデンス・ベースを盲信する現代社会において現実化しているまさにその現象であると言っても過言ではないだろう。その意味で，社会科学の碩学であるキャンベルの論稿（Campbell, 1979）に先見の明があったということは決して言い過ぎではない。

また，社会科学の中で「キャンベルの法則」ほど注目されているわけではないが，キャンベル（1979）では，もう一つ重要な指摘もしている。それは，「過大評価の罠」(overadvocacy trap)である（Campbell，1979，p.84）。通常，社会実験を行うような大規模に資金をかけた「社会問題」というのは，アメリカの人種問題や貧困問題に代表されるように，そもそも慢性的に解決できない問題であったり，通常であればうまく機能している標準的な機関がそもそも失敗している問題である，という指摘（Campbell，1979，p.84）は正鵠を得ている。加えて，議会は，政治的パフォーマンスとして，社会問題を解決することよりも，それ自体はなかなか解決できないため（解決できるのであれば，アメリカ社会でもこの世の中からも貧困や人種問題はとうの昔に無くなっているだろう），ただ行動を起こしていることを重視するために，実証研究自体には予算が不足がちになるのが常であり，それが完璧には実証されにくいがために，本当に価値のあるプログラムであったとしても誇張された結果になるか，もしくは，意外に低い効果しか生じていないようにしか評価されない事態を生じさせるという（Campbell，1979，p.84）。このことは，統計的因果を厳密に実証するということとはまた別の，測定評価において構造的に生じる異なる政治的な次元の問題として考えることが適当なのであろう。

3．教育テストにおける「キャンベルの法則」の事例

加えて，キャンベル（1979）は，教育についても，い

わゆる「キャンベルの法則」が成り立つことを述べており（Campbell，1979，p.85），現代の教育学者からもしばしば言及されている（例えば，後述する Koretz，2008・2017）。キャンベル（1979）は，学力テストそのものは，「一般的な学校の達成度を示す貴重な指標となる」（Campbell，1979，p.85）と認めながらも，「テストの点数が教育プロセスの目標となってしまうと，教育状況の指標としての価値が失われ，教育プロセスが望ましくない方向に歪められてしまう」（Campbell，1979，p.85）と言及している。その具体的な事例として，請負業者が生徒の学力テストの成績に応じて報酬を得る場合，報酬欲しさ故に、生徒に学力テストの答えを教えてしまったというスキャンダル事例を紹介している（Campbell, 1979, p.85）。また，より大きな効果に見せかけようと，事前テストを意図的に点数が低くなるように実施したり，事前テストで低い点を取った生徒を調査や集計から除外することで，平均点を意図的にあげること事例なども紹介している（Campbell，1979，p.85）。教育でしばしば生じるこれらの事態は，生徒自体にとって、数値が指し示しているような，例えば，学力が向上したなどの事態が存在せず、実質には何も進歩がないことを意味する。

ハーバード大学の教育測定学者であるダニエル・T・コレツ（Daniel T. Koretz）は、「キャンベルの法則」が教育テストにおいて現れる現象を「スコア・インフレーション」という用語を使って説明している。コレツ（2017）によれば，「スコア・インフレーション」とは，「学習が真に改善しているよりもスコアの伸びが大きいこと」（Koretz，2017，p.46）であり，本当は「点数を上げる効果がないにも関わらず，教室で行われていることを改善するどころか，むしろ悪化させているものもある」(Koretz，2017，p.46)という意味で，「キャンベルの法則」の要件を満たす事例として考えられている。

「キャンベルの法則」のわかりやすい例として，コレツ（2008）では，飛行機における飛行時間の統計の事例を挙げている（Koretz，2008，pp.237-8）。実際には，出発時間に遅延しているにも関わらず，定時に到着したり，定時運行しているかのように宣伝されているが，実際にはさほど目的地に早く着いたようには思えない経験的事例を挙げ，その実，飛行時間そのものには大幅に早くなるなどの変化がないことを考えれば，ダイヤ上の飛行時間が以前より長く設定されたがために，定時運行が見掛け上守られていることになっているに過ぎないことを指摘している。これは，「実際には、『定時』の意味を再定義することで，改善されたように錯覚させた」（Koretz，2008，p.240）事例であり，飛行時間そのものには何も変化がないことから，ルールを変えただけに過ぎず，実際には何も本質的に変化していないことを意味する。

実際には何も本質的に変化がないにも関わらず，見掛け上の数値が変化する現象について，コレツ（2017）は長年に及ぶ自身の教育テスト研究の経験に基づいて，興味深い事例を紹介している。それは，コレツ（2017）が，「鋸の歯ような模様（the sort of sawtooth pattern）」（Koretz，2017，p.59）と名づけている現象であり，測定の道具であるテストを変更することによって生じる無意味な波形を指す（図 1）。図 1 は，コレツが提示した図であり，縦軸の数値は，学年換算点（Grade Equivalents : GE）と呼び，学校教育の任意の時点での典型的なスコアを 1 学年を 10 ヶ月とした学校教育の月数を表す指標である。3.7 という数値は，37 ヶ月であり，全国の小学生 3 年生が 7 ヶ月過ぎた時点での得点の中央値を指す。この指標は，「スコア・インフレーション」を可視化するために，コレツ独自に開発した指標である。

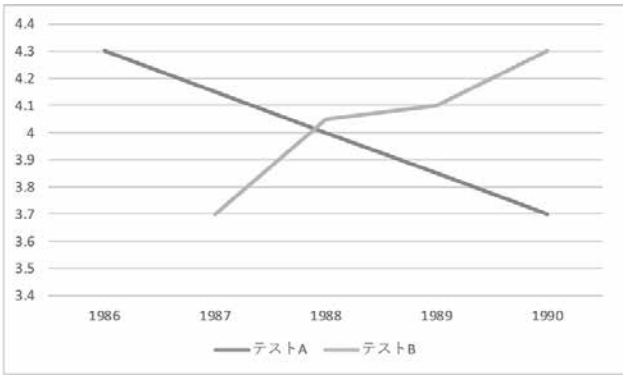


図 1. テスト A からテスト B の変更によって生じるスコアの高低（Koretz，2017，p.60 の figure 5.2 より筆者作成）

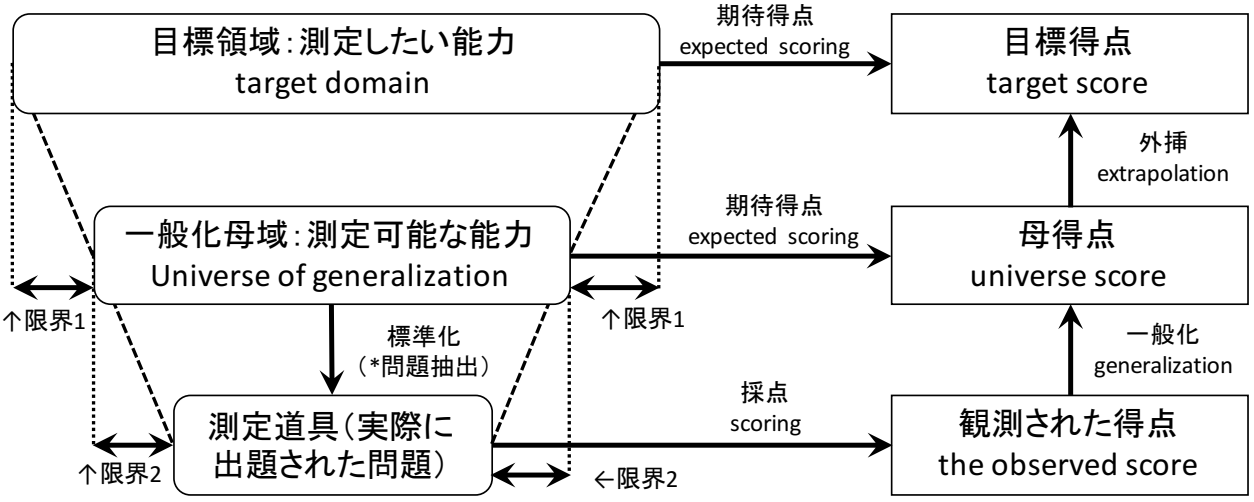
これによると，ある地区でテスト A からテスト B に変更した際，GE のスコアは，元々4.3 であったものから，3.7 に急落することを図 1 は表している。だが，これは急に当該地区の生徒の学力が，0.6 つまり半年分低下したことを本当に意味するのだろうか。また，3 年経過すると，GE スコアは元の 4.3 に戻っている。このことは，当該地

区の学力が本当に回復したことを意味するのであろうか。コレツ（2017）ではははっきりとは述べられてはいないものの，実質的には何も変化がない事例として挙げられているのは自明のことであろう。

もちろん，数値単独では，何が急落や回復の原因であるのかを説明することはできない。これを悲観的に，その地区の学力が低下したと喧伝することも可能であろうし，「奇跡」と称して，学力の回復や向上を喧伝することも，社会で見られる現象としてはありうるのかもしれない。一方，コレツ（2017）は，1990 年代に急激に州テストの得点を向上させ，白人とアフリカ系およびラテン系の生徒の学力格差が劇的に改善されたと喧伝された「テキサスの奇跡」に言及している。そして，この「テキサスの奇跡」が，後に大統領となるジョージ・W・ブッシュが全米で注目されるきっかけとなり，選挙キャンペーンの中で重要な役割を果たしたことを指摘している（Koretz，2017，p.67）。加えて指摘するのは，この州テストの「外部妥当性」の問題であり，全国学力調査であるテキサス州の NEAP のスコアにはほとんど変化がなかったことである（Koretz，2017，p.67）。州テストと全国学力調査である NAEP が同時に大幅なスコアアップがあるのであれば，傍証という意味も含めて，生徒になんらかの変化が起こっている可能性を否定しにくくなるだろう。その意味で，同一測定領域の異なるテストの結果が一致していることが望ましく，これは教育測定論で元来，「収束的妥当性」（convergent validity）や「収束的証拠」（convergent evidence）と呼ばれてきたものである。

コレツ（2008，2017）では，このような「スコア・インフレーション」が生起する可能性について，2 つの説明をしている。1 つ目の説明は，教育指導の「ゼロ・サムゲーム」によるものである。教育指導をする時間が総計としては変化しないとすれば，ある領域や項目についての指導に力を入れた場合，それ以外の領域や項目についての指導する時間が減る（Koretz, 2008, pp.237-8, 2017, p.63）のは言うまでもなく，そうしたメカニズムがスコアに表れているとする説明である。ただし，こうした事態が成り立つのは，「テストを作成するために使用されたサンプルリングが予測可能である」（Koretz，2017，p.62）状態にある場合とし，その予測可能な内容としては，テストの問題内容，傾向，出題方法，生徒の解答パターン（誤解の類型を含む）を挙げている（Koretz, 2017, p.62）。

2 つ目の説明は，テストバッテリーが測定したい母領



テストの限界1*：「測定したい能力」と「測定可能な能力」とのズレ
テストの限界2*：「測定可能な能力」の範囲と実際に出題した問題が問える範囲のズレ

採点(scoring)：観測されたパフォーマンスが評価され、観測得点が生じる
一般化(generalization)：観測された得点が、目標得点を推定するために一般化される。
含意(implocatiion)：推定された目標得点を言葉による記述に翻訳する。
外挿*：未知の目標得点を既知の観測された得点から一般化された母得点から推定する。
(*は筆者による説明追記)

図 2. 測定母領域とテスト項目と得点との関係

（木村，2016，p.101 の図 1 より筆者作成）

域のサンプルであるという考え方である（Koretz, 2017, p.58）。これは，木村（2016）で示した図2で説明する方がわかりやすいだろう。つまり，テストには，測定したい能力である「目標領域」があるが，それについての測定可能な能力は限定され，それは「一般化母域」と呼ぶ。更に，テストには満遍なくその領域から出題できればいいが，通常，それには限界があり，実際に出題されているテスト項目は，「一般化母域」からの抽出されたサンプルであると考える。テストとはそういう想定する測定領域と測定可能な測定領域，そして実際に測定する測定領域とがズレた構造を孕むものであり，測りたいものが全て測れるものと対応しているわけでもなく，観測されたスコアには，ある種の限定がかかっている状態であると理解しやすいだろう。その意味で，コレツ（2017）が強調するのは，仮にテストAとテストBのテストとしての「目標領域」は同じであったとしても，測定道具としてのテスト項目が違う可能性があり，テストAの回答に慣れていた生徒が，テストBでスコアが当初うまく伸びない事態は十分起こりうる事態であり，その原因は，「一般化母域」から異なるサンプルとして抽出された項目であったとして説明可能である，ということになる。

つまり，コレツ（2008，2017）では，テストスコアが「鋸の歯ような模様（the sort of sawtooth pattern）」（Koretz, 2017, p.59）になる原因について，教育指導とテスト構造という異なる２つの次元から指摘したわけだが，実は，その両者には繋がりもあり，テストへの慣れは，ある特定の教育指導の結果と捉えることも可能であろう。

また，コレツ（2017）は，「スコア・インフレーション」について，「多くの州で読解力を調査した研究では，無料または割引価格のランチを受けられる生徒の方が，他の生徒よりもインフレーションの発生率が高いことがわかっている」（Koretz, 2017, p.68）とし，その原因として，「リソースの不足，経験の浅い教職員，教員の離職率の高さ，生徒の転出率の高さ，模範となる成績の良い生徒の少なさ，補助的な支援を提供できる保護者の少なさ，保護者からの学業成績へのプレッシャーの低さ」といった「障害に直面すると，教師はスコアをあげるための近道を探す動機付けが強くなる」（Koretz, 2017, p.68）とも説明している。テストのスコアに過度に説明責任を持たせることによって，本来，読み取れるはずのテストスコアからの情報が歪んでしまうことは誰にも益をもたらさない。コレツ（2008）は，「スコアが誇張されると，それに基づいて人々が判断する最も重要な結論の多くが間違ったものとなり，その結果，生徒や，時には教師が苦しむこととなる」（Koretz, 2008, p.233）と述べ，教育テスト分野における「キャンベルの法則」について警鐘を鳴らしている。肝心なのは，「テストの準備によって得られたスコアの上昇が，生徒の達成度の意味ある上昇を本当に表しているのかである」（Koretz, 2008, p.258）ということであり，「テストの特定の項目を教えることを，『Teaching the Test』といい（明らかに悪い），テストが表すはずのスキルに焦点を当てることを『Teaching to the Test』と呼ぶ（おそらく良い）」（Koretz, 2008, p.251）と区別して論じている。

教育テストにおける「キャンベルの法則」について，コレツ（2008）は，「基準に基づいたテストの時代には，整合性（alignment）が政策の要となる」（Koretz, 2008, p.253）と述べているが，この言葉にコレツの主張が収斂されている。つまり，実際に生起している現実とテスト・スコアに整合性があって初めて，「（ここでは「テスト」の）数値」が社会にとって真の意味を持ちうることを強調したのである。

4. まとめ－測定評価を巡る構造的陥穽からの示唆

総括すると，キャンベル（1979）であれ，コレツ（2008, 2017）であれ，巷に溢れるような安易な測定批判を行いたいのではなく，測定結果が実際に社会で活用されるための条件について言及した，と言う方が正確であろう。

測定評価を巡る問題は，冒頭でも言及したように，要約すれば，数量化され視覚化されたものを受け入れる，歴史の中で時間をかけて築かれてきた精神的素地の上で，測定結果を受け入れるだけでなく，それに執着しやすい，あるいは，「科学的」な結果として真に受けやすいという事態を生じさせることに始まる。加えて，測定値には解釈という価値づけの問題が含まれているが，それ以前に測定値の算出は統計・解析技術によるものであり，その技術自身が社会からブラックボックス化することも含めて，測定評価を巡る構造的陥穽とまとめることができる。「キャンベルの法則」は，こうした測定評価を巡る現代社会の見えづらい構造的陥穽に1つの注釈を加えるもの

であると評価することができるだろう。

「キャンベルの法則」が生じる裏側で生起しているのは測定結果と現実とのズレであり，コレツ（2017）はその整合性こそ重要なものとみなしていた。キャンベル（1979）でも指摘されていたように，教育測定論の立場からすれば，それは測定によって得られた数値の「外的妥当性」の問題である。1 つの解析結果は，ある特定の解釈を生み出し得るが，別の同様の条件で取られたデータからも同じことが言えるのか，と言う意味では，心理学で現在問題とされている再現性の問題と近似している。ただ，社会実験の場合，多くは，いわゆる「同様の条件で取られたデータ」であると言う条件を満たすことが困難であり，しばしば，得られたデータやそこから導き出された解釈の独壇場になり，データを取得した関係者以外に，「有無を言わさない」構造に陥りやすいことは言うまでもない。

理想的に言えば，二種類以上の解析結果による検証可能性が必要とされるが，相矛盾する解析結果としての数値が存在する場合，同等の条件で得られたものであるかが，まずは問題とされる。もし，仮に同等の条件で得られたものであるとするならば，次は，その数値の解釈可能性の問題となるだろうが，現実にはどちらも理想的な測定条件を満たしていない，ある意味限界のあるデータからの解釈が相矛盾していることが多いと考えられる。数値を錦の御旗として無条件に受容するのではなく，相矛盾する数値が出現した際に，そのどちらの数値が正しいのか，そしてそれが誰の責によって判断されるのかという意味では，即座に政治的な問題となることが多く，社会的意思決定を伴う場合，統計とは全く異なる観点（例えば，より社会的に望ましいとか，それ自体が政治的な観点）での意思決定がなされることが多いだろう（例えば，木村 2021 を参照のこと）。

その意味では，得られた数値が正しいかは，実験デザインの問題であったり，ひいては統計的因果推論の議論になったりするか，あるいは，社会調査の観点から言えば，サンプルの代表性の問題にもなりうる。だが，それらはいわゆるデータ解析の厳密さと解釈の妥当さを論じるものでもある。その意味では，これらの観点からの吟味は，より白黒つきやすいという意味で議論が比較的容易なものであるとも言える。だが，「キャンベルの法則」で問題としているのは，統計的な厳密さや上述した意思決定における政治的次元の問題とは，また別の種類の政治的次元として測定評価がもたらす影響である。「キャンベルの法則」は，人物選考において関連する別の要因や変数がその評価自体に影響を与えやすいハロー効果の話と，純粋な計量測定の問題ではない別の要因や変数が最終結果に関連するという意味では，類似構図をもつものなのかもしれない。「キャンベルの法則」が我々に指し示してくれるのは，「測定にまつわる問題」と，「測定からもたらされる問題」という二種類の問題が存在するということではないだろうか。これこそが，測定評価に関する問題構図の根源に存在する。これら二つの問題は，どちらが大事とか，どちらがより優れているのかという話ではなく，両者を混同することなく峻別して，測定評価を社会で正しく議論することが必要となることを教えてくれる。

＜注＞

(1) 例えば，別名として，「コブラ効果」（Sierbert，2001）や「ラット効果」（Vann，2003）が存在するほどである。ヴァン（2003）では，フランスの植民地であった 19 世紀末のベトナムにおいて，ペストを媒介するとして衛生面から駆除の対象であったラットに対して，尻尾を持ってくることに対して懸賞金をかけたところ，ラットが減るどころか，街中には尻尾のないラットが溢れ，ラットを飼育する人まで現れ，1902 年にはペストが街で流行したことを紹介している。

(2) ミュラー（2018=2020）も，警察の事例を著書の中で取り上げている。

＜引用・参考文献＞

Campbell, D.T., 1979, “Assessing the Impact of Planned Social Change”, in *Evaluation and Program Planning*, 2, pp.67-90.

Crosby, A.W., 1997, *The Measurement of Reality—Quantification and Western Society, 1250-1600*, Cambridge, Cambridge University Press,（=2003，アルフレッド・W・クロスビー，小沢千重子訳『数量化革命—ヨーロッパ覇権をもたらした世界観の誕生』紀伊国屋書店）.

Gould, S.J., 1981=1996, *The Mismeasure of Man—The definitive refutation to the argument of The Bell Curve,*

revised and expanded with a new introduction, New York, W・W・Norton & Company, （=1989=1998，スティーブン・J・グールド，鈴木善次・森脇靖子訳『人間の測り間違い―差別の科学史 増補改訂版』河出書房新社）.

木村拓也，2016，人物重視の大学入試は「妥当」か？--大学入試改革論議のテスト理論的理解，『教育と医学』64 巻 2 号(通巻第 752 号)，、pp.30-38.

木村拓也，2021，米国大学入学者選抜における大規模標準化テスト SAT/ACT からの離脱決定の論理構造―カリフォルニア大学における標準化テスト・タスクフォース（STTF）報告書の分析，『九州教育学会紀要』48，pp.25-32.

Jencks, C., 1972, *Inequality: a reassessment of the effect of family and schooling in America*, New York, Harper & Row,（=1978，橋爪貞雄・高木正太郎訳，『不平等―学業成績を左右するものは何か』黎明書房）.

Koretz, D., 2008, *Measuring Up: What Educational Testing Really Tells Us*, Massachusetts and London, Harvard University Press, pp.1-275.

Koretz, D., 2017, *The Testing Charade: Pretending to Schools Better*, Chicago and London, The University of Chicago Press, pp.1-275.

Muller, J.Z., 2018, *The Tyranny of Metrics*, New Jersey, Princeton University Press,（=2019，ジェリー・Z・ミュラー，松本裕訳『測りすぎ―なぜパフォーマンス評価は失敗するのか？』みすず書房）.

Poter, T. M., 1995: *Trust in Number—The Pursuit of Objectivity in Science and Public Life*, New Jersey, Princeton University Press,（=2013，T. M. ポーター，藤垣裕子訳『数値と客観性―科学と社会における信頼の獲得』みすず書房.

Siebert, H., 2001, *Der Kobra-Effekt. Wie man Irrwege der Wirtschaftspolitik vermeidet*. Munich, Deutsche Verlags-Anstalt.

Vann, M. G., 2003, “Of Rats, Rice, and Race: The Great Hanoi Rat Massacre, an Episode in French Colonial History”， in *French Colonial History*, 4, pp.191–203.