特集

国際教育開発協力のこれまで・これから

【トピック編：分析手法の深化】

計量経済学的教育評価の作法

伊藤 成朗

教育は人間が成長し、社会の一員として活躍するために欠かせない過程である。このため、どのような教育が望ましいのか不断に評価して、改善していくことが望まれる。

子どもの発達を促す教育にはさまざまな要素がある。教育の場所から考えても、学校、家庭、補習施設、企業など、学習機会は幅広い。それぞれにおいてどのような評価の仕方が相応しいかは、教育の目的や文脈に応じて考えられるべきである。本稿では、評価対象に最もなりやすい学校教育を例に、計量経済学的な評価の仕方や作法を考えてみたい[1]。

一般に、教育の効果を計測するには、結果指標、母集団、評価手法、標本数を選ばねばならない。具体的な手順をみるために、以下では、アメリカのチャーター・スクール（charter school 以下「CS」）の効果についてのアメリカ教育省報告書（参考文献③、以下、「報告書」）を参考にしながら、計量経済学的評価の作法について考えていく。CSとは、教育到達目標を憲章にすることで、政府補助金を得ながら運営される私立校である。学費が私立校よりも低い一方で、論者によっては、公立校よりも質の高い教育を提供する（参考文献④）、営利組織なので教育効果は低い（参考文献⑩）、などと評価が分かれている。報告書では、CSに通学すると公立校通学よりもどれだけ共通試験の点数が良くなるかを検討している。

●結果指標

結果指標は評価で知りたいことを最も適切に表す指標を選ぶ。成績に関心があれば試験点数、試験合格率、進級率、自己規律であれば出欠遅刻率や宿題提出率、知的好奇心であれば自由課題の結果などである。報告書では州テストの国語と数学の点数を結果指標として選んでいる[2]。

途上国でも就学率が高まっているため、教育の成果・結果指標には、学習内容を示す成績が採用されはじめている。成績は知的能力発達の尺度として適切であり、自己規律なども反映する。数値化できる指標なので、変化を観察しやすい。しかし、成績が学校の予算や評判に影響する場合、成績だけを結果指標として取り上げると、学校や教師の関心が成績に偏りかねない[3]。このため、可能であれば複数の成果を表す複数の指標を選ぶことが望ましい。

成績以外の指標としては、成績とは無関係の指標を選ぶと良い。同じ対象から得た複数の結果指標は相互に相関しやすいことから、似た指標が複数あっても情報が増えないためである。例えば、五教科が性質の異なる科目であっても、能力の高い生徒ほど多くの科目に秀でる可能性があるので、政策効果の検定を三科目から五科目に増やしても新しい知見が無いかもしれない[4]。それよりも、チーム競技成績や創意工夫による掃除時間の短縮化など、成績と無関係で客観的に計測可能な数値の方が総合的な学習成果として参考になる。

報告書では、学校ごとに効果の有無を検定し、同一科目の検定結果が各CS間で相関している可能性に配慮しているが、国語と数学の試験点数間の相関は考慮していない。よって、報告書の読者は、個別の教科に関する検定結果よりも、学校全体を見渡した集計的な検定に信頼を置くべきである。

●母集団

母集団（population）とは、特定の特徴を共有する事象の集合である。教育評価の文脈では、評価対象になっている生徒たちと同じ特徴を持つ生徒の集団を指す。異なる対象でも、母集団が同じならば

同じ政策効果を期待できるが、母集団が異なれば同じ政策でも効果は異なるであろう。評価結果を他の対象に応用するためには、母集団を明確に定義すべきである[5]。

報告書では、定員を超える出願者があったCSを対象にし、そのうち、評価研究への参加に同意し、かつ、前年度の成績を提出できなかった一部の私立学校出身者以外の生徒を対象にしている。つまり、CSが出願者を多く集める地域で、主に公立学校出身の生徒である。CS出願者が多いということは、公立校がより不人気の地域なので、公教育予算の少ない貧しい地域の可能性が高い。報告書によれば、生徒一人あたり予算は評価対象のCSで八〇三〇ドルに対して、対象外のCSは八七一〇ドルである。よって、CSの学習効果は、平均よりも貧しい地域の公立校出身生徒に対する効果と解釈できる。

●評価手法

インパクト評価は、「政策が実施された状態」と「政策が実施されなかった状態」を比較して効果を測定するのが原則である。しかし、同じ対象に政策が実施された状態と実施されない状態を同時に観察することはできない。このため、評価では、政策の影響を受けなかった対象（「統御群」control group）を実施されなかった状態とみなす必要がある。その際には統御群が政策の影響を受けた対象（「処置群」treated group）と同じ母集団に属すると仮定するのだが、その仮定が現実的でなければ、評価の信頼性は低い。

例えば、一五歳生徒が対象の教育政策を考えよう。この政策の成績への効果を評価するときに、一四歳生徒を統御群として使うためには、両者の特徴は共通している、と仮定する必要がある。この仮定が成り立つためには、一四歳と政策実施前の一五歳の成績の分布が同じで、一四歳に政策の効果が及んではならない[6]。もしも、一五歳で学ぶ数学が一四歳の数学よりも難解であり、一五歳になると全員が成績を下げる場合、一四歳生徒の成績分布は一五歳生徒の成績分布に比して高めになる。「政策が実施されなかった状態の一五歳生徒」の成績としては一四歳生徒の成績は高すぎるために、一五歳生徒の成績が政策によって高まったとしても、政策効果は過少に推計される。また、一五歳対象の政策が一四歳にも影響を及ぼせば、一四歳生徒の成績は「政策が実施されなかった状態の一五歳生徒」の成績を表さない。このように、統御群とみなすための仮定の現実性を考えれば、評価の信頼性が決まってくる。仮定の現実性は評価の文脈に依存するので、その都度、評価者が判断しなくてはならない。

最も信頼性の高い評価方法はランダム化比較試験（randomized controlled trial：RCT）である。RCTでは、評価対象をランダムに処置群と統御群に割り振り、両群の結果指標平均値の差を平均処置効果（average treatment effect：ATE）とみなす。RCTで統御群を「処置群が政策の影響を受けなかった状態」とみなすために必要な仮定は、ランダム化が正しく行われたこと、偶然に異質な群が二つできないほど標本規模が大きいこと、統御群に政策の影響が及ばないことである。これらは評価者が統御可能なので、仮定が現実的である可能性は高く、評価の信頼性も高いことが多い[7]。

RCTにも短所はある。第一に、実験であることが分かると、実験関係者の行動を変える可能性である。対象者が実験下にあることを知ると行動を変える可能性（ホウソーン Hawthorn 効果やジョン・ヘンリー John Henry 効果）、実験実施者が実験下にあることを意識して細心の注意を払い、実施内容が実際の政策とは異なる内容になる可能性（研究バイアス research bias）などである。自然実験（natural experiment）は実験関係者が実験になっていることを気付かないので、この短所はないが、対象を選べないという短所がある。第二に、論理的に、倫理的に操作できない内容がある。例えば、実兄の存在が子どもの発達に与える影響は、兄を今から産むことは論理的にできないし、兄のいる家庭からランダムに兄を奪うことも倫理的ではない。第三に、稀な事象は観察機会が少ないので、標本規模が相対的に小さいRCTに不向きである[8]。

報告書では、くじを使った入学選考過程をRCTとみなしている。出願者多数の場合、CSではくじを使って入学者を決めることが多い。報告書が評価のためにランダム化（くじ）を導入したのではなく、既に実施されているくじ抽選制度を利用したのである。既存のランダム化の利用は、評価の

手間と費用が低く抑えられるので有益である[9]。報告書では両群が同質か検討するために、CS通学前の両群の特徴が統計的に有意に異なるか検定している。検定の結果、前年度の数学の成績が処置群の方が高かった以外、成績や家庭環境に有意な差は無く、両群は総じて同質と判断している[10]。

●標本サイズ

標本規模が小さいと政策の効果を検知できない可能性がある。効果が検知できないほどの小標本の場合、評価を実施することは倫理的ではない、と判断されることもある。これを避けるためにも、評価者は評価を実施する前に検定力分析（参考文献⑪）をして標本サイズを決める。

学校対象の評価における標本の選び方には特徴がある。学校対象の評価では母集団すべての学校を対象にすると費用がかさむので、一部の学校を選択する。学校選択後は、一校あたりで調査する生徒数が増えても費用はかさまないので、全生徒を調査することが多い。この標本抽出設計は一段クラスター抽出法（one-stage cluster sampling）という。クラスターとは集団のことであるが、ここでは学校がクラスターである。

標本サイズが同じ場合、クラスター抽出法は単純無作為抽出法（simple random sampling：SRS）よりも費用を低く抑えられる一方で、偏った情報を得る危険がある。例えば、一校あたり一〇人×一〇校＝一〇〇人の標本を得たとしても、各校内の生徒が似ていたら、一校あたり生徒数を増やしても新しい情報を得にくい。極端な仮想例を挙げると、各校内の生徒が全く同じ複製であれば、一〇人×一〇校＝一〇〇人から情報を得ても一〇人分の情報しか得られない。このように、クラスター内相関係数（intracluster correlation coefficient）が高い場合、標本サイズが大きくても実質的には小標本でしかない。クラスター抽出法にすると費用を節約できるものの、一定の効果を検知するために必要な最小標本サイズをSRSより増やさなければならない[11]。これに対応して、統計的検定に用いる標準誤差も、クラスター内相関を許容したクラスター頑健標準誤差（cluster-robust standard error）を使うことが推奨されている[12]。

報告書では二五校から合計一二〇〇人の生徒を抽出することを想定し、五％有意水準の下、確率八〇％でインパクトが点数の標準偏差の一四％以上なら統計的に有意と検定できることを示している。ただし、この検定力分析はクラスター内相関に言及しておらず、SRSを想定して計算された可能性がある。そうであれば標本サイズ計算としては不十分である。クラスター内相関を考慮していたとしても、読者にはその内容が分からないため、報告書として配慮が不十分である。とくに、報告書がCSの効果を見出していないのは、クラスター内相関を想定せずに計算した標本サイズが小さすぎたからかもしれない。もしもそうであれば、CSの効果が認められないという結論をどこまで受け入れるべきか、直截な判断はできない。

●結果の導出

五％有意水準で評価デザインを設計すると、真の効果がゼロであったとしても、平均すると一〇〇回に五回は帰無仮説が棄却できずに効果ありという結果を得る。このため、推計方法を微妙に変えながら有意な結果が出るまで何度も検定（「データ・マイニング」）すれば、有意な結果を出すことができる。これを避けるため、データをみる前に検定する仮説や推計方法を決める分析前計画（pre-analysis plan）を公開することも推奨されている。データ・マイニングを許すと、効果ありという研究ばかりが世に出る出版バイアス（publication bias）を起こすためである[13]。

評価で最も大事なのは、何を知りたいか、高い信頼性の下に知ることができるか、という二つの問いである。先行研究に当たれば、すでに同じ目的の評価があって新たな評価は不要かもしれない。想定している教育政策の効果を知りたい理由（CSで学ぶと成績が上がるか）を突き詰めると、結果指標や効果発現過程の解明に必要な情報（教員や級友の質、クラス規模）の示唆が得られる。想定する効果発現過程に応じて、適切な評価設計（くじによるRCT）も導き出されるであろう。その評価設計での処置群と統御群が同じ母集団に属する、とみなすために必要な仮定が非現実性ならば、無理に評価をするよりも、信頼性のより高い設計で評価を将来にする方が良い、という判断に落ち着くかもしれない。得たい知識と比して評

価の信頼性が十分に高いのか、評価者は常に自問自答しなければならない。

（いとう　せいろう／アジア経済研究所　在ステレンボッシュ海外調査員）

《注》

(1)計量経済学的な評価は定量的評価である。評価には定性的な手法もある。定性的な評価は、政策が成果を上げる過程を細かく示し、利点や問題点を析出する一方で、政策と成果の因果関係の論拠は曖昧である。定量的な手法は、政策と成果の因果関係を示す一方で、政策が効果を発揮する過程の分析は粗雑である。このように、定性的評価と定量的評価は一長一短があり、それぞれの評価を相互補完的に設計すべきである。もちろん、それぞれの手法には長所があるにしても、個別の評価においてその長所が十分に発揮されているという保証はない。

(2)州テストは州毎に試験問題が異なるので相互比較が難しい。報告書では試験点数をz値に変換している。z値とは各値xをその期待値μと標準偏差σによって標準化したものである。

$$z=\frac{x-\mu}{\sigma}$$

z値はxと期待値との差を共通のばらつき尺度（標準偏差単位）に換算するので、異なる分布間でも、州間でも、比較可能である。

(3)試験点数が重視されると、試験点数を上げるテクニックを教えることが優先され、学校で学ぶべき教科の理解、知的好奇心、級友との友好関係構築スキル、多様性への寛容などは蔑ろにされかねない。極端な場合、教師が試験問題の内容を教えたり、採点に手心を加えるなどの不正を引き起こすこともある（参考文献⑤）。

(4)複数の結果指標を使って効果の有無を検定しても、複数の独立した検定とみなすことはできない。複数指標を検定する際には、検定統計量間の相関を考慮しなければならない（参考文献①②⑥⑦⑨）。

(5)他の対象に政策効果の評価が当てはまる場合、その評価には外的妥当性（external validity）があるという。とある母集団について効果を歪みなく測っている場合、その評価には内的妥当性（internal validity）があるという。

(6)正確には、その他の共変数を所与とした条件付き分布が同じである必要がある。政策効果が処置群内に留まるという仮定は、SUTVA（stable unit treatement value）と呼ばれる仮定から導出される。SUTVAは標本の結果指標が他標本に割り当てられた処置に依存しない、という仮定である。

(7)一方で、差の差（difference in differences）、傾向値（propensity score）、統御関数（control function）、合成統御法（synthetic control method）などの推計方法は、処置群と統御群が同じ母集団であるとみなすための仮定が非現実的である場合が多い。

(8)他にもRCTとその他の方法に共通の短所がある。第一に、政策結果によって追加的な処置が発生する場合である。例えば、早期児童発達政策によって処置群の子どもが愛らしくなり、周囲の人間から追加的な世話を引き寄せるとしよう。すると、結果指標が政策効果以上の変化を示すが、効果を政策と追加的な世話に分離することはできない。第二に、研究や政策実施に同意しない対象者は参加せず、同意者のみの効果しか計測できないため、非同意者への効果が分からないことである。RCTにも参加同意という自己選抜過程があるために、母集団のなかの同意者に関する効果しか計測できない。この意味でRCTも内的妥当性に限界がある。同意と不同意がどのように発生するか予測できなければ、他の対象への適用可能性（外的妥当性）も限られる。

(9)評価者は出願者多数のCSのうち、入学者を決めるのにくじを使っている学校を選び、そのなかから研究参加に同意した親と生徒を標本としている。母集団はくじ採用CS入学希望者（くじ不同意者も含む）のうちの調査同意者、処置群は母集団でくじに当たってCSに通う生徒、統御群は母集団でくじに外れて公立校に通う生徒である。くじによるCS入学機会提供の効果は、くじ不同意者の効果（＝0）とくじ同意者の効果の加重平均である。これは「政策意図に基づく効果」（intention-to-treat effect：ITT）と呼ばれる。くじ不同意者がCSに通学した際の効果がゼロよりも大きい場合、ITTはATEよりも小さくなる。

⑽偶然に両群に差異が発生しても、回帰式に差異のある変数（前年度数学の成績）を加えればその影響を考慮できるので、ＲＣＴを採用している限り、効果を歪みなく計測でき評価の信頼性も高い。

⑾この増加率を設計効果（design effect：Ｄｅｆｆ）という。各クラスターで同じM人、クラスター数Nが大きく$NM-1$が$M(N-1)$に近い場合には、設計効果は以下の式で与えられる。

$$Deff = (M-1)\rho$$

ここでρはクラスター内相関係数である。よって、ρが高いほど（＝学校内の生徒が似通っているほど）、一校あたりの生徒数Mが大きいほど、設計効果は大きい。仮に、校内の相関係数が〇・二五、一校あたりの標本数が四一の場合、設計効果は一〇なので、ＳＲＳよりも一〇倍多い標本を必要とする。ただし、この場合でも、クラスター抽出法の方が調査費用は安いかもしれない。

⑿クラスター頑健標準誤差は通常の標準誤差よりも大きくなるので、効果なしという（帰無仮説を棄却しない）場合が増える。クラスター内相関への配慮は一九九〇年代に入って経済学で浸透してきた分析作法である。これに対し、一部の計量経済学者からは、クラスター頑健標準誤差は過剰棄却傾向がある（oversized）という批判もある。

⒀しかし、データをみる前には変数の分布の様子すら分からず、不適切な推計方法を選んでしまう可能性もある。これも含めて分析前計画を書けば良いのだが、すべての可能性を網羅することはできないので、データ覗き見（data snooping）によって推計方法改善のヒントを得られる余地が残る。どれだけの覗き見を許すのかは議論が続いている（参考文献⑧）。

《参考文献》

① Anderson, Michael L. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association*. 103 (484): 2008. pp.1481-1495.

② Benjamini, Y., and Y. Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society*, Series B (Methodological). 57 (1): 1995. pp.289–300.

③ Gleason, P., M. Clark, C. C. Tuttle, and E. Dwoyer. "The Evaluation of Charter School Impacts: Final Report (NCEE 2010-4029)." Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. 2010.

④ Hoxby, Caroline M. "Does competition among public schools benefit students and taxpayers?" *American Economic Review*. 2000. pp.1209–1238.

⑤ Jacob, Brian A., and Steven D. Levitt. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *The Quarterly Journal of Economics*. 118 (3): 2003. pp.843–877.

⑥ Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz. "Experimental Analysis of Neighborhood Effects." *Econometrica*. 75 (1): 2007. pp.83–119.

⑦ O'Brien, Peter C. "Procedures for Comparing Samples with Multiple Endpoints." *Biometrics*. 40 (4): 1984. p.1079.

⑧ Romano, Joseph P., Azeem M. Shaikh, and Michael Wolf. "Formalized Data Snooping Based on Generalized Error Rates." *Econometric Theory*. 24: 2008. pp.404–447.

⑨ Westfall, Peter H., and Stanley S. Young. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley Series in Probability and Statistics. Wiley-Interscience. 1993.

⑩ Wiggin, Addison. "Charter School Gravy Train Runs Express To Fat City." *Forbes*. September 10, 2013. Available: http://onforb.es/19FeQ2q [Last accessed: October 2, 2014].

⑪高野久紀「実践開発経済学 Vol.2 ランダム化比較試験（フィールド実験）検出力分析」『経済セミナー』二〇一四年八・九月号。