経営論集　第30巻第1号　2020年　23–38頁

エビデンスに基づいた教育

Evidence Based Education

小　松　香　爾

〈論文要旨〉

信念は経験の蓄積から生じる。生徒・学生の経験は、学校教育を通じて獲得されたものが多い。社会人の経験は、組織で働くことから獲得されたものが多く、教員の経験は、教育実践で獲得されたものが多い。

個人的な経験の蓄積から生じた信念は、書籍・論文等から得た知見によって修正されたり、変更されたりする。しかし、そのような知見が、科学的な根拠に基づいている保証はない。

教育分野においては、信念のみにしか基づかない根拠薄弱な言論が多い。そして、そうならざるを得ない理由がある。本論文では、教育工学に内在する脆弱性を論じることにより、教育分野の言論が根拠薄弱である理由を示す。

〈abstract〉

Beliefs arise from the accumulation of experience. Many of the students' experiences have been gained through school education. Many of the experiences of working people have been gained from working in an organization. Many of the teachers' experiences have been gained through educational practice.

Beliefs arising from the accumulation of personal experience are modified or changed by the knowledge gained from books, papers, etc. However, there is no guarantee that the knowledges are scientifically based.

In the field of education, there are many weakly grounded speeches that are based only on beliefs. And there is compelling reasons. In this paper, I present the reasons why the grounds must be weakened by discussing the inherent vulnerabilities of educational technology.

〈キーワード〉

科学、根拠、教育工学

2020年10月26日受理

1. 科学であるための必要条件

科学の応用が技術である。技術の誕生は科学的発見より後であり、数百年後、数千年後のことかもしれない。人類滅亡の方が先の可能性もある。しかし、応用できる可能性がなければ、科学的な研究ではない。応用できる可能性があれば、人類社会の進歩に貢献できる可能性がある。

1888年に電波を発見したヘルツは電波の有用性を認識できていなかった。現代社会は、電磁波の利用なしには成り立たない。電波なくしては、無線放送・無線通信ができないからである。

ヘルツは、物理実験で、マクスウェルによって構築された「電流と磁場の相互作用によって電磁波が生じる」という電磁波理論の正しさを証明した。電磁波理論の証明により、電波の存在を実証した。

科学的な研究であるための必要条件は応用可能性の他にも３つある。

（1）応用可能である。すなわち、現在あるいは将来において、現実世界の問題を解決できる可能性がある
（2）普遍性と再現性の存在
（3）因果関係の存在
（4）根拠（データ）の存在

科学とは、狭義には自然科学である。自然科学の代表は物理学である。広義には、科学は、社会科学と人文科学も包含する。教育学は、人文科学に分類されるが、教育学における研究対象は人間の行動である。心理学や社会学などと共に、教育学は、行動科学として分類されることもある。教育工学は、教育学の一分野である。教育工学の研究では、統計学という応用数学が用いられる。統計的手法で、実験やアンケートから得られたデータを分析するのが定石である。

1.1　普遍性と再現性

いつ、誰が、どこで、何度試行しても、プロトコルさえ守られれば、同じ結果になるのが普遍性と再現性である。

iPS細胞には普遍性と再現性があったが、STAP細胞には、どちらもなかった。ただし、iPS細胞は技術的にES（Embryonic Stem）細胞より、技術的に作成が困難であり、また再生医学的にも使いにくい。人工的に作成された細胞であるため、癌化する可能性が大きくなるからである。一方、ES細胞の作成には受精卵が必要なので、ES細胞には倫理的な問題がある。また、ES細胞は、再生医療を受ける人自身の細胞から作られた細胞ではない。拒絶反応が出る可能性が大きくなる。

医学は、自然ではなく、人体が研究対象であるため、この種の解決困難な問題が生じる。特に、受精卵を再生医療で使用して良いか否か？という問題は、倫理学・哲学・社会学で取り扱

われるべきである。科学では可否の判断ができない。

STAP細胞の論文は、追試によって普遍性・再現性が確認されなかった。STAP細胞の論文に限らず、当初は普遍性・再現性が高いとされていた研究が、他の研究者による追試によって、否定されることがある。

ニュートン力学は、光速に近い運動をするモノに対しては成立しない。ただし、静止しているモノにかかる重力は、地球の重心からの距離が同じという条件さえ揃えられれば、どこでも等しい。万有引力の法則は、普遍性と再現性が極めて高いといえる。

1.2　因果関係

原因があって結果がある。その関係が因果関係である。物体に力が加わった時に、加速度が生じて物体の速度が変化する。このとき、力が原因、加速度の発生が結果である。力と加速度の因果関係は、以下のニュートンの第2法則で表される。

$$m\vec{\alpha} = \vec{F}$$

上記の式で m は物体の質量、$\vec{\alpha}$ は加速度、$\vec{F}$ は力である。 $\vec{F}$ は素粒子間に働く、強い力、電磁気力、弱い力、重力の4つの力から生じることが定説である。4つの力の伝達粒子は、それぞれグルーオン、フォトン、ウィークボゾン、グラビトンである。そのうち、グラビトンだけが2020年10月9日現在では、未発見である。したがって、「重力がなににによって伝わっているか？」という問いについては、現在の科学は答えられない。

ただし、ある物体の質量をM、もう1つの物体の質量をm、両者の距離をr、万有引力定数をGとしたとき、以下の万有引力の法則が成り立っている。

$$F = G\frac{Mm}{r^2}$$

万有引力の法則は、万物（ただし光速では移動していない）に対して成り立っている法則である。太陽と地球、地球と月、人間と人間、電子と原子核の間など、2つの物体があれば、上式通りに働く。この万有引力により、リンゴが木から離れると地球との間の引力が働いて下に落ちるという再現性が担保される。

リンゴの質量をm、地球の質量をM、重力加速度をg、物体と地球の重心との距離をRとすると、以下の式が成り立つ。

$$mg = G\frac{Mm}{R^2}$$

上の式の両辺を質量の値mで割ることにより、重力加速度gは以下の式から求まる。

$$g = G\frac{M}{R^2}$$

上式では、m が変化しても、重力加速度 g は変化しない。つまり、重力加速度は、リンゴだけでなく人間や素粒子など地球上のすべての物体で同じである。時間経過があっても、質量不変であれば変化しない。重力には、因果関係および普遍性も担保されている。

1.3　根拠

科学的な研究の根拠は、現象に関する実験や観測・調査のデータである。

物体と物体に重力加速度が生じる原因の万有引力の間には、伝達粒子グラビトンという素粒子が存在すると仮定されている。そう仮定しても他の現象・法則とは矛盾しないというだけで、現時点では仮説にすぎない。したがって、グラビトンは存在しないかもしれない。重力は自然界に存在する4つの力のうち最も弱く、現在の科学技術では、存在を実証することはできない。しかし、万有引力を伝達する物質の挙動が解明されていなくても、現実に観測される現象との矛盾は生じていない。

科学的な判断を下す際には、実験や観測のデータが必要である。因果関係があり、それと矛盾しない各種データがあるため、万有引力には科学的な根拠がある。したがって、万有引力は定説となっている。万有引力の法則によって、「惑星の円運動の原因となる力は太陽の引力である」と説明できるようになり、地動説も揺るぎのない定説となった。

科学的な根拠がない場合は、科学では結論を出せない。しかし、科学な根拠があっても、科学で出せる結論は暫定的な結論である。結論を出せたとしても、未来永劫にわたって、結論の正しさが保証されるわけではない。

科学的に真とされていたことが、新たな実験や観察・調査から得られたデータ（科学的な根拠）によって、偽になることもある。科学の進歩で生まれた技術によって、それまで不可能であった実験や観察・調査が可能になる。天動説は地動説によって否定された。ニュートン力学は、相対性理論と量子論によって、一部が否定された。

一般相対性理論では、質量を持った物体が動くと、時空のゆがみが光速で伝わる重力波が生じるとされていた。しかし、1916年当時は、検出のために必要な技術がなかった。一般相対性理論を構築したアインシュタインは、重力波の検出は不可能であると予測した。100年の時がたち、2015年に、重力波が検出された。科学の進歩によって開発されたレーザーマイケルソン干渉計によってである。

レーザーマイケルソン干渉計では、ビームスプリッタで分けられた2つのレーザー光の干渉の変化を観察することによって、重力波を検出する。重力波が干渉計に到達すると、重力波の影響により2つのレーザー光の進む空間が伸縮する。空間が伸縮すると、レーザー光の干渉にずれが生じる。光の干渉のずれが、重力波の到達を示す。2015年に検出された重力波、太陽

質量35.4のブラックホールと太陽質量29.8のブラックホールの衝突の際に発生したことによって発生したものであった。

重力波の検出により、電磁波では捉えられない宇宙で起きる事象を、観察できるようになった。これまでは観察できなかった現象を観察できるようになるため、新しい定説が生まれたり、既存の定説が否定されたりする可能性が大きくなる。

2　科学の限界

仮説の検証が科学的な研究の定石である。しかし、検証につかうツール（技術、数学）も人間が考えたものであり、能力に限界がある。

2.1　仮説の価値

科学の一般的な研究において、研究者は最初に仮説を立てる。次に、なんらかの実験か観察・調査を行う。最後に、実験や観察・調査から得られたデータ（科学的根拠）に基づいて、最初に立てた仮説を検証する。検証した結果が、社会に効用をもたらす可能性があれば、学会発表や論文誌への投稿などで公開される。その一連のプロセスは、同一の研究者によって行われるとは限らない。

通常は、仮説が真であると判明した場合に研究が公開される。世の中に公開される研究結果はそのようなものである。しかし、公開されていなくても研究結果に価値がある場合もある。仮説が真でなくても、科学的な根拠が示されていれば、価値のある研究といえる。「当該仮説は偽である」可能性が大きくなるからである。少なくとも、その時点の技術では、真と示すことができない可能性は大きくなる。

2.2　プロセスとしての科学

科学はプロセスにすぎない。現在の科学では、科学的に実証できないことが多数存在する。例えば、人体に及ぼす放射線の悪影響である。人間の細胞に及ぼす放射線の影響は確実に存在する。高線量の放射線が、DNA鎖に当たれば、DNA損傷が生じる。しかし、数ミリシーベルト程度の低線量の放射線を浴びた場合に、発がん率などが上昇するかどうかについて、科学では判断できない。現時点では、因果関係を実証するデータ（科学的根拠）が存在しないからである。「上昇する」あるいは「上昇しない」というデータの、いずれもが存在しない。

研究対象が自然現象であれば、新たな測定装置が発明され、観測や実験を重ねることで科学的根拠が作られる可能性は大きい。しかし、科学的には可能であっても、倫理的には不可能な実験は多い。特に、人体に、不可逆的な影響を与える可能性が大きい実験では、倫理的な問題が発生する。人体に及ぼす放射線の影響に関する実験は、その典型例である。

放射線の人体への影響は、確定的影響と確率的影響に分けられる。確定的影響は、高線量の

放射線に被爆することによって生じる急性放射線障害である。確定的影響が発生する最小線量である閾線量が存在し、閾線量を超える被爆線量の増加と共に、脱毛などの発症率は100%に達するまで増加する。確率的影響は、比較的低線量の放射線に被爆した後に起きる、生殖細胞の突然変異やがんである。原爆投下による被爆者の疫学的研究では、被爆者の子孫に遺伝的影響は認められなかったとされている。しかし、生殖細胞の変異は、確率的影響である。原爆投下は日本における2例しかない。1945年の2例に対する疫学的研究のみをもって、「放射線の遺伝的影響がない」という判断はできない。

自然放射線から人体がうける放射線量の平均は年間約2.4ミリシーベルトである。自然放射線以外に、レントゲンやCTスキャンなどの医療用放射線からうける放射線量の平均は、年間約2.4ミリシーベルトである。低線量の放射線に関して、「100ミリシーベルト未満の被爆に確率的影響はない」とする閾値仮説と、「ゼロより大きい放射線量は線量に比例して確率的影響が発生する確率が上がる」とする直線閾値なし（LNT：Linear-Non-Threshold）仮説がある。これら2つの仮説に関して、双方とも真である可能性は0%である。双方とも偽である可能性も0%である。しかし、現在の科学では、どちらが真でどちらが偽かを証明することはできない。

国際放射線防護委員会（ICRP：International Commission on Radiological Protection）は、直線閾値なし仮説を採用し、LNTモデルという線量反応モデルに基づいた勧告を行っている。しかし、確率的影響が確率的であることと、LNTモデルには閾値が存在しないという性質とが、低線量放射線の安全性判定を不可能にしている。

ICRPは、2007年に、公衆被爆の個人線量限度を1ミリシーベルトと勧告した。緊急時被爆状況における公衆被爆は、20ミリシーベルトから100ミリシーベルトと幅を持たせている。

2011年の福島第一原子力発電所の炉心融解事故後、原子力安全委員会は、事故発生から1年間の積算線量が20ミリシーベルトに達する恐れがある地域を計画的避難区域と指定した。20ミリシーベルトという値はLNT仮説に基づくものである。仮説は仮説にすぎず、実験や観察・調査で実証されたわけではない。実証されれば、仮説は定説となり、仮説ではなくなる。20ミリシーベルトを閾値とした計画的非難区域の指定が妥当であったかどうかは、当時も、2020年10月9日現在も、科学的には判断できない。それにも関わらず、政治的には判断する必要が生じた。現実世界で生じる事象は、科学の進歩を待ってくれない。しかし、将来的には、科学的に判断できるようになる可能性はある。科学は現在進行形で進歩しているからである。

2.3　科学で取り扱わない仮説

微量な放射線の人体に対する悪影響とは異なり、地動説は、現在の科学（物理学・天文学）では、数々の実験・観察結果から、ほぼ確実に真である。地動説に限らず、現在、定説となっている自然現象に関する命題のほとんどは、今後も真でありつづける[1]。現在、否定されている

[1] 数学というツールの強力さによる。公理の健全性と無矛盾性により、全命題の真偽を判定できる。

自然現象に関する命題のほとんどは、今後も偽でありつづける。しかし、ラッセルの「世界五分前仮説」は、科学的には真偽を判定できない。

世界五分前仮説では、「世界は5分前にできたのではないこと、および5分以上前の過去が存在することを証明することは不可能」とされている。実験・観察の結果も、人間の認知に依存することは間違いない。科学的には、世界五分前仮説の否定は不可能である。

世界五分前仮説が真であると仮定しても、人間による未来の世界への介入の結果は変化しない。人間の介入行動による影響が起きないならば、問題解決はできない。世界五分前仮説は真である可能性もあり、実証される可能性もないとはいえないが、現実世界で応用できる可能性がない。したがって、哲学的には意味のある仮説[2]であるが、科学では考慮する価値がない仮説と見なされる。

3　エビデンスの信頼性

エビデンスというキーワードは、疫学分野で初めて使われた。疫学は自然科学に分類されるが、人間の疾病とその要因との関連を解明する学問分野である。研究対象は人間の集団であり、物理学と比較すれば、頑健な法則が少ない。自然を対象とする科学の法則は頑健である。科学的根拠のある現象しか、法則として認められないからである。

人間の集団にも法則は存在する。定説となった仮説は法則と呼べる。しかし、自然の法則ほど頑健ではない。したがって、エビデンスの有無よりも、エビデンスの信頼性が重要である。エビデンスがあれば、仮説が正しいわけではない。エビデンスがあっても、エビデンスの信頼性が低ければ、仮説が真である確率は小さい。仮説が真である確率が小さければ、当該仮説が定説として定着する可能性も小さい。

1989年に社会を騒がせた常温核融合には、エビデンスがあった。しかし、再現性と普遍性に乏しくエビデンスの信頼は低かった。現時点でも信頼性の高いエビデンスはなく、「常温でも核融合現象が起きてエネルギーが抽出できる」という常温核融合仮説は、定説とはなっていない。しかし、仮説が実証され、真であれば、エネルギー問題の解決につながる。

現在の原子力発電はリスクが大きい。リスクよりも効用の方が大きいとみなされている国が多いため、原子力発電所は世界の多くの国に存在する。しかし、核分裂は発熱が極めて大きいため、原理的にリスクをゼロに近づけることはできない。事故が起きれば、チェルノブイリ原発事故や福島原発事故のような大惨事になる。常温核融合仮説が実証されれば、人類にとっての効用は極めて大きい。したがって、仮説が真である可能性が低いとされている現在でも、多数の研究者がエビデンスを探索中である。ただし、エビデンスは、発見し、発表すればよいと

[2] 哲学は、思考プロセスを研究する学問である。思考実験を行うこと自体が目的であるので、哲学において意味のない仮説は存在しない。科学において、実験は目的ではなく、仮説を検証するための手段である。

いうものではない。信頼性の高い、つまり、再現性と普遍性が高いエビデンスでなければ、人類の進歩に役立つ研究となる可能性は小さい。

教育学では、1990年代後半から、教育実践や教育政策はエビデンスに基づくべきであるという考え方が広まった。OECDのCERI（Center for Educational Research and Innovation）では、2004年にEvidence-based Policy Research in Educationというプロジェクトがスタートした。EBE(Evidence-Based Education)と呼ばれているが、教育現場、特に、日本の教育現場には浸透していない。

教育とは、つきつめれば脳内のニューラルネットワークを再構成する行為である。脳の機能はごく一部分しか、科学的に解明されていない。脳内で、シナプスと神経伝達物質による電気信号の伝達が起きることは既知であり、実証されている。しかし、電気信号の伝達で、なぜ人間は感じたり考えたりできるのかは全く分かっていない。したがって、信頼性の高いエビデンスが得られる観察や実験を行えない。

エビデンスに基づく医療であるEBM(Evidence-Based Medicine)は、EBEより早い時期から提唱されていた。EBMは、1991年に提唱された。(Guyatt 1991）1990年代中盤から日本の医療現場に導入された。ニュートンの3法則が1687年に発見されてから、300年以上たって、EBMが提唱され、定着したことになる。

人間を対象とする観察から、普遍性と再現性のある法則を発見することは困難である。人間を対象とする実験から、因果関係の存在を実証することが困難だからである。また、それらの観察結果と実験結果が矛盾しないことを検証するのも困難である。医学でも、特に精神医学の分野では、信頼性の高いエビデンスは得られにくく、EBMが確立されにくい。脳の実験・観察が、倫理的・技術的に困難だからである。

EBMは精神科を除く医療現場に浸透した。エビデンスとしての信頼性が高い研究は、標準治療として医療現場に反映される。エビデンスには、エビデンス・ピラミッドとよばれる信頼性のレベルが存在する。同じ信頼性レベルにある研究でも、バイアスの排除状況や、サンプルサイズの大小によって、エビデンスの信頼性は変化する。エビデンス・ピラミッドは、細部の研究デザインを考慮していない、大まかな基準にすぎない。

3.1　専門家の意見

EBMのエビデンスのうち、最も信頼性が低いのは、専門家の意見である。専門家の意見は、専門家の経験から帰納的に得られた結論を含有している。全く信頼性がないわけではない。しかし、公開されたデータに基づかない意見は、信頼性の高いエビデンスとはいえない。

EBMに限らず、科学においては、論より証拠による実証が原則である。意見が巧みに構成されて理論としてまとまっていても、実証されていなければ、エビデンスとしての信頼性は低い。

文科省が打ち出す教育政策は、中央教育審議会などに属する専門家（教育学の研究者とは限らない）の意見を集約したものである。したがって、教育政策が基づくエビデンスも専門家の

意見になりがちである。EBEが浸透していないため、教育政策は、信頼性の高いエビデンスには基づいていない。

3.2　動物実験・基礎実験

動物実験や培養した細胞を試験管内などで使う基礎実験から得られるデータ、それらのデータから導き出される理論も、専門家の意見と同様に、エビデンスの信頼性が低い。動物と人間では、生理学的な機能は大きく異なる。動物で起きる現象と人体に起きる現象が同じあるいは似ているとは限らない。

試験管内で人為的に起こせる現象は、人体で起きる現象のほんの一部に過ぎない。一部で成り立つことが、全体でも成り立つとは限らない。人体は複雑度が高い複雑系であり、全身の健康状態などは分析が困難である。

動物実験・基礎実験で得られるデータ自体は科学的な根拠である。「細胞レベルで有効ならば人体にも有効な可能性がある」や「動物で有効ならば人間にも有効な可能性がある」などの仮説を立てることはできる。しかし、部分的に真であることが、全体で真であるということを担保するわけではない。したがって、仮説を立てるための研究の中でも、エビデンスとしての信頼性は低い。

教育工学では、学力の一部しか、評価項目として設定できない。学力とは「暗記力」と「考える力」と「意志力」という3つの大まかな因子に分解できそうではある。しかし、暗記力はペーパーテストで評価できているとしても、後者2つがペーパーテスト・レポートで評価できているのか、あるいはペーパーテスト・レポートで評価すべきものか、文科省からも見解がでていない。学習指導要領のタイトルは「生きる力」など、誰からも批判されようがないものにシフトしてきている。変化の大きい現代社会を生きるために「生きる力」が必要というのはトートロジーである。情報量が増えていない。

資本主義社会では、「生きる力」は「学力」ではなく「カネを稼ぐ力」であるという信念を、文科省官僚の主流派が持っているように見受けられる。諮問機関での会議に、産業界のエスタブリッシュメント層が呼ばれているからである。しかし、「カネを稼ぐ力」は、その時の環境、すなわち、所属する組織、国家体制、国際情勢、為替レート、科学の発展度等、さまざまな要因に大きく依存する。それらの外部要因と独立して「カネを稼ぐ力」は定義できない。「カネを稼ぐ力」が社会の経時変化に対して、一定であることはありえない。

人間は神ではない。神でないならば、社会の経時変化を予測することはできない。したがって、「カネを稼ぐ力」の持続時間も予測できない。現代社会の変化のスピードは激しく、「カネを稼ぐ力」をどう定義しようと、普遍性と再現性を持たせるのは不可能である。さらに、「学力」と「カネを稼ぐ力」との相関・因果関係も不明である。高偏差値の大学の卒業生の平均年収が高いという統計データは存在する。しかし、平均年収からは、企業の新卒採用における学歴フィルタ等の選考基準という、選択バイアスを排除できない。

教育分野においては、アウトカム、特に長期的なアウトカムの測定が困難である。文科省は、「大学受験を変えなければ高等学校における教育は変わらない」という信念のもと、2020年度実施の大学入試を改革した。受験制度を変えてから、結果が出るまでには時間がかかる。受験制度の改革の「結果らしきもの」がでるのは、それらの受験に合格し、教育を受け、社会に出て働いて10年以上働いてからである。すなわち、20年後あたりにようやく「アウトカムらしきもの」がでる。しかし、それくらいの時間が経てば、社会は激変する。大学入試改革が実施されたときには稼げる人材だった人間が、20年後にも稼いでいる人材である保証はない。

3.3　症例報告

症例報告は、1人～数人の患者の治療経過や治療結果を、詳細に記述する研究である。記述的研究とも呼ばれる。症例報告から、仮説を立てることはできるが、エビデンスとしての信頼性は低い。症例報告は、稀な症例を報告したものがほとんどである。患者が典型的な経過をたどらなかった場合や、医師が教科書的な治療とは違った治療を行った場合などで、実施される。したがって、観察対象の集団が小さい。対象となる集団が小さいので、実験ができない。症例報告は、レポートとしての価値はあるが、研究デザインが存在しない。

教育工学でも、特定の教育者が特定の集団に対して一時的に行った教育実践報告は、多数存在する。研究資源、すなわち時間、資金、モチベーション、能力などの乏しさや、実験デザインの拙さなどにより、エビデンスとしての信頼性は、EBMにおける症例報告程度しかないものもある。

3.4　横断研究

横断研究は観察研究の一種である。ある時点における、ある要因の有無と疾病の有無を調査し、要因と疾患の関連性を評価する研究である。横断研究では追跡調査が行われないため、多くの対象者を低コストかつ短期間で調査できるメリットがある。しかし、横断研究は、ある時点での調査にすぎず、要因と疾患の前後関係を調査できない。原理的に、相関関係の有無は実証できても、因果関係の有無は実証できない。

物理学の法則は、因果関係を記述している。工学も、物理学の応用であるため、一部、因果関係が解明されていない部分は残るものの、大部分は、物理学で解明された因果関係に基づいた研究が行われている。しかし、教育工学には、基礎となる法則が存在しない。あるいは存在するのかもしれないが、実証されていない。研究者の直感・感覚・主観・経験に基づく仮説しか存在しない。

学力の柱と著者がみなしている「考える力」と、読解力との間に相関関係があるという仮定は、「考える力」の定義次第ではあるが、真となる可能性が大きい。しかし、仮定が真であっても、「読解力がある⇒考える力がある」あるいは「考える力がある⇒読解力がある」という因果関係があるのか、「考える力」と「読解力」が同じものであるのか、それとも「考える力」

と「読解力」は別物であり、相関関係があるだけなのか、解明されていない。また、将来的に解明される見込みもない。「考える力」も「読解力」も定義できていないからである。

現在、日本には「読解力」に関するRST（Reading Skill Test）と呼ばれるテストが存在する。（新井 2019）RSTでは、読解力は7つの因子に分解される。係り受け解析、照応解決、同義文判定、推論、イメージ同定、辞書的具体例同定、理数的具体例同定である。その7つの因子の説明文の中には、「把握」、「認識」、「判定」、「判断」、「理解」という語句が使われている。これらの語句は「考える力」と、相関か因果か同値のいずれの関係にあるのかについて、説明はされていないし、見解も存在しない。巷に仮説としては存在するかもしれないが、仮説が定説とはなっていない。7つの因子の説明文は便宜的定義であり、包括的な定義とはいえない。

3.5　症例対象研究

症例対照研究も観察研究の一種である。時間的な前後関係を調査するため、横断研究よりエビデンスとしての信頼性が高い。症例対照研究では、疾患がある集団である症例群と、疾患のない集団である対象群を設定し、特定の要因Aにさらされた過去の状況を調査する。状況調査は、医療記録や面接・アンケートで実施される。横断研究と異なり、因果関係をある程度は捉えられる。しかし、面接・アンケートを使った症例対照研究では、想起バイアスのリスクが大きい。一般的に、症例群の方が、要因Aにさらされた記憶を思い出しやすい。想起バイアスの存在によって、本来、症状の原因ではないものが、原因とされる可能性が生じる。また、症例群と対照群を設定する際にも、選択バイアスが生じやすい。症例群と対照群は、様々な属性が同じ１つの母集団から抽出されるのが理想である。しかし、一般的に、症例群は入院患者の集団から、対照群は入院患者以外の患者の集団から選ばれる。性別、年齢、生活習慣などの条件が、できるだけ症例群と一致するような対照群を設定することで、選択バイアスがかかるリスクを小さくすることはできる。しかし、症例対象研究は、過去に遡って調査する後ろ向き研究である。後ろ向き研究では、過去に遡って交絡因子を調査することが困難である。交絡要因とは、原因と疾患の両方に影響を及ぼし、真の因果関係とは異なる結果をもたらす要因のことである。後ろ向き研究は、ある時点から未来に向かって観察する前向き研究よりもエビデンスとしての信頼性は低い。

3.6　コホート研究

コホート研究は観察研究の一種である。観察研究では最もエビデンスの信頼性が高い。コホート研究では、ある時点で観察する横断研究や、ある時点から過去に遡って調査する症例対照研究と異なり、ある時点から未来へ前向きに観察する。コホート研究では、母集団を、特定の要因Aに曝露された集団と、されていない集団に分けて長期間観察し、要因Aの影響を比較調査する。注目する疾患の罹患率が低い場合、大きな集団を長期間観察する必要が生じる。時間とコストがかかることがデメリットである。また、交絡要因の存在にも注意する必要がある。

年齢と性別は、多くの場合、交絡要因となる。年齢や性別の属性が、要因Ａと疾病の両方に影響する交絡因子となることが既知の場合には、それらの属性を揃えた母集団を研究対象にすることによって、間違った因果関係の導出を防ぐことができる。しかし、未知の交絡要因の影響は排除することができない。交絡要因が既知であっても、交絡因子を排除すれば、集団のサイズが小さくなる。観察対象が小集団になればなるほど、エビデンスとしての信頼性は低くなる。したがって、コホート研究では、同一地域の同一年代を母集団とすることが多い。

3.7　前後比較研究

観察研究は、仮説を立てる、あるいは立てられた仮説を分析するために実施される。それに対して、介入研究は仮説を検証するために実施される。仮説を立てたり分析したりするだけの研究業績では、自然科学分野のノーベル賞候補にはならない。ノーベル賞受賞の必要条件は、理論が実証されていることである。自然科学分野における理論を記述するための強力なツールである数学における研究業績は、ノーベル賞の対象ではない。

前後比較試験は、介入研究の一種である。介入研究では、研究対象に投薬・治療などの介入を行い、効果を検証する。前後比較試験は、介入研究の中ではエビデンスとしての信頼性が一番低い。ある程度以上の大きさの集団に対して、ある要因Ａを加えたとき、変化Ｂを観察でれば、要因Ａが変化Ｂの原因となったという仮説を立てられる。しかし、要因Ａを加えない対照群が設定されていなければ、生じた結果を比較できない。前後比較試験は、症例研究と同様に、稀な疾患等の理由で、対象者を集められない場合に実施される。

3.8　非ランダム化比較試験

非ランダム化比較試験も介入研究の一種である。前後比較試験よりエビデンス・レベルが高い。非ランダム化比較試験では、実験対象を母集団からサンプルとして抽出し、サンプルを投薬や治療などを行う介入群と、それらを行わない対照群に割り付け、介入群に対する介入の効果を対照群と比較して評価する。「非ランダム」とは、介入群と対照群への割り付けがランダムではなく、人為的に行われるという意味である。例えば、診察時間、主治医、病院などの違いで、介入群と対照群を割り付けた場合、両群の属性が偏りやすい。割り付けが人為的だと、アウトカムに対して選択バイアスがかかる可能性が大きくなる。したがって、エビデンスとしての信頼性は、3.9節で記述するランダム化比較試験より低くなる。

実験において、介入行動以外の人為的な行動が少ないほど、エビデンスとしての信頼性は高くなる。教育学・教育工学においても、対照群を設定して、介入群に対して、効果があると仮定する教育を実践し、対照群と介入群で教育効果の違いを観察するような研究は存在する。しかし、教育実践の場合、実施者・実施場所・実施時間などの諸条件を揃えることが不可能であることが多い。また、教育効果の違いは、過去に受けた教育や家庭環境にも依存する。それらの条件を揃えることは不可能である。

教育の研究においては、「何をもって教育効果（アウトカム）とするのか」に関して、明確なコンセンサスが存在しない。研究論文では、学生に対してアンケート調査をしているものが多い。アンケートには学生の主観が入る。医療の臨床試験における5年生存率や腫瘍縮小率などのデータと比較すると、教育実践のアンケートのデータは客観性で劣る。学生本人が教育効果を感じていても、実際には効果がない場合、あるいは学生本人が教育効果を感じていなくても、実際には効果がある場合がある。

臨床試験において、期待される治療効果（アウトカム）はエンドポイントと呼ばれる。最終的なエンドポイントはトゥルー・エンドポント（true endpoint）と呼ばれ、死亡率・発症率の低下やQOLの向上である。しかし、前者は時間的な問題があり、後者は時間的な問題に加えて客観性の問題がある。いずれも、観測して、客観的・普遍的に求めることが困難である。したがって、代替的なエンドポイントであるサロゲート・エンドポイント(surrogate endpoint)として、腫瘍の大きさや血圧や血糖値など、短期間で、数値の変化を観測可能な評価項目を設定することが多い。

教育工学では、臨床試験におけるトゥルー・エンドポイントは存在せず、サロゲート・エンドポイントに相当する評価項目の妥当性は低い。テストの結果として、算出される偏差値は数学的に厳密な定義があり、母集団の範囲内では客観的な評価項目である。しかし、偏差値には普遍性がない。特定の母集団に対して実施した、特定のテストにおける、集団中の相対的な評価しかできない。さらに、テスト自体が、学力を測る目的で適切とは限らない。したがって、偏差値＝学力ではない。偏差値には、再現性はともかく、普遍性がない。偏差値の上昇は、学力の上昇の信頼性の高いエビデンスにはならない。

少なくとも文科省による学力の定義が存在しない。学ぶための力なのか、学んだ結果の力なのか、その両方であるのか。両方であるとしたら、両者は分離可能であるのか。仮説としては存在するかもしれないが、定説は存在しない。定義されていないものに対しては、直接的には科学的研究を行うことが不可能である。幽霊は科学的に研究できない。学力は、すくなくとも現時点では、幽霊同様に定義がないものである。

3.9　ランダム化比較試験

ランダム化比較試験（RCT：Randomized Controlled Trial）は、介入研究の中で最もエビデンスの信頼性が高い。介入群と対照群への割り付けが、コンピュータで生成した乱数などを用いて機械的に行われるからである。介入群と対照群へ、被験者をランダムに割り付けることにより、母集団からサンプル抽出される際に生じる選択バイアスを消すことができ、交絡因子の偏在を防ぐことができる。つまり、非人為的な割り付けにより、介入群と対象群で、アウトカムに影響を及ぼすような要因が偏らないことを担保できる。ただし、ランダム化比較試験でも、母集団のサイズが大きくなければ、交絡要因が偏在しやすくなる。したがって、母集団のサイズが小さいRCTは、エビデンスとしての信頼性は低くなる。

非ランダム化比較試験とRCTの共通点は、介入群と対照群を設定した後に、介入群へは投薬などで介入し、対象群には偽薬などでプラセボを生じさせるのみで介入はせず、エンドポイント到達後に、両群に生じたアウトカムの差を比較調査することである。

教育工学には、プラセボ効果を生じさせる偽薬に相当するものが存在しない。仮説としては存在する可能性はあるが、定説にはなっていない。実験的な教育を受ける生徒・学生には、工夫された従来と異なる教育を受けているというバイアスがかかる。同等のバイアスを、従来型の教育を受ける生徒・学生にもかけられる偽薬のような教育があれば、アウトカムには影響を及ぼさない。しかし、学校教育においては、プラセボ効果を生じさせる偽薬のような教育は不可能に近い。幼稚園における教育であれば可能性はあるが、年齢が進むにつれ不可能に近づき、高等教育においては不可能である。

3.10　システマティックレビュー・メタアナリシス

ランダム化比較試験の論文を集めるのがシステマティックレビューで、ランダム化比較試験の論文のアウトカムを信頼性の高さで重みづけするのがメタアナリシスである。ランダム化比較試験自体のエビデンスとしての信頼性が高い。システマティックレビュー・メタアナリシスは、信頼性の高いエビデンスを集積する研究である。したがって、エビデンスの信頼性が最も高い。

システマティックレビューでは、ランダム化比較試験の論文を網羅的に調査し、研究デザインをチェックすることにより、ランダム化比較試験の中でも信頼性の高い論文のデータのみを抽出する。メタアナリシスでは、母集団の大きさなどで、各データ、すなわちリスク差（Risk Difference）、リスク比（Risk Ratio）、オッズ比（Odds Ratio）などのアウトカムを、母集団のサイズで重みづけして統合する。システマティックレビュー・メタアナリシスのエビデンスとしての信頼性は高く、学術論文に引用されやすい。

教育工学においても、レビュー論文は存在する。しかし、プラセボ効果を生じさせる教育が成り立たないという点において、アウトカムにバイアスがかからない比較試験ができない。また、教員に対しても、アウトカムを期待できる教育を実施しているのか、それとも従来型の教育を実施しているのかを識別不能にすることができない。教育においては、臨床試験における2重盲検化ができない。人が人に対して教育するかぎり、原理的にできない。したがって、ランダム化比較試験らしき教育実践を実施し、論文を書いたとしても、エビデンスとしての信頼性は低い。信頼性の低いエビデンスを集めても、信頼性は高くならない。

4　おわりに

医療分野でも、EBMは1990年代まで普及しなかった。対象が人間であるゆえに、実験や観測が自然現象と比べて困難だからである。教育分野では、実験・観測の実施の対象が人間であり、多くの場合、実験・観察を実施するのも人間である。

エビデンスとしての信頼性が高い研究であるための必要条件は3つ挙げられる。

（1）実験をしていること
（2）母集団が大規模であること
（3）バイアスが取り除かれていること

教育分野においては（2）と（3）が困難である。特に（3）に関しては、教育では、被験者の盲検化が不可能であるため、バイアスの完全な除去は不可能である。

教育分野には、学力の定義がないため、教育効果の定義も直接的にはできない。したがって、（1）の実験においても、間接的に教育効果に関係しそうなアウトカムを測定・評価することしかできない。教育工学の研究者は、仮説レベルのアウトカムを実証するために実験することしかできない。エビデンスに基づく教育は困難である。

SDG'sでも、17の持続可能な開発目標のうちの1つとして「4 QUALITY EDUCATION」が謳われている。質が低いより、質が高い方が良いに決まっている。しかし、なにをもって「質が高い教育」とするのかについて、コンセンサスが存在しない。コンセンサスが存在するとしても、定説となっていない。したがって、「課題が山積み」という教員は多いが、ではその課題が何なのかを具体的に述べる教員は少ない。

定義されていない幽霊のようなものに対しては、誰でも、もっともらしいことが言える。しかし、発言に、根拠がなく、具体的な課題解決法にも結びつかないのであれば、発言は意見ではなく、雑談である。雑談を楽しむ目的ならば問題ないが、教育の質をあげるという目的に対しては、時間的な損失になる可能性が大きい。であるならば、個々の教員は、個々の教員の教育上の信念（エビデンスはない）に基づき、教育内容の充実を目指すべきではなかろうか？少なくとも、教育内容と教育方法を教員に任せる方針が、生徒・学生の価値観・コンピテンシーの多様性を許容・醸造する可能性は大きいと信じている。しかし、エビデンスはなく、著者個人の信念にすぎない。

参考文献

Guyatt G.H.（1991），　Evidence-Based Medicine (editorial），ACP Journal Club，Annals of Internal Medicine，vol. 114, suppl.2
新井紀子（2019），AIに負けない子どもを育てる，東洋経済新報社