波及効果に関するリサーチデザインの検討：
大学入試改革をエビデンスに基づき議論するために

須藤　爽

1. はじめに

本稿の目的は、波及効果（washback effect）に関する研究が政策的示唆に寄与するために、いかなるリサーチデザインが求められるかについて論じることである。2020 年度の大学入試改革の目玉ともいえた、英語試験への「民間試験導入」に関する是非は世間を大きく騒がせた。最終的には 2019 年 11 月 1 日にその延期が発表されたものの、中止はせず、2024 年度からその導入再開が予定されている。

民間試験導入は、「テストを 4 技能評価にすることで、高校での学びを変える」、すなわち、民間試験導入による波及効果を引き起こすことを一つの目的としている。波及効果に関する研究には長年の蓄積があり、今日でも世界各国で幅広く扱われる研究テーマである。しかし、そのほとんどがテストの妥当性の検証を目的としており、政策的観点からみるとエビデンスの質はかなり低い（寺沢, 2019）。したがって、民間試験導入に関する議論は、エビデンスの軽視どころかそもそもエビデンスが存在しないため、関係者の主観的な憶測に基づき進められているのが現状だ。この点を踏まえ、波及効果に関するエビデンスの質を向上させるためにどのようなリサーチデザインが必要かを論じる。

本稿の構成は次のとおりである。まず、第 2 節において民間試験導入の経緯とその目的を振り返る。そして第 3 節以降で、波及効果のリサーチデザインのあり方について検討し、波及効果に関する研究のエビデンスの質を高め、政策的示唆につなげるためにはどのようなリサーチデザインが必要かを考察する。

2. 民間試験導入の経緯と目的

2.1. 民間外部試験の導入が延期されるまでの経緯

2020 年度の大学入試改革は世間の受験生・受験関係者を大きく振り回した。特に英語民間試験導入案については、2017 年 7 月に文部科学省（以下「文科省」と言う）は「大学入学共通テスト実施方針」の中で、民間外部試験 （以下「民間試験」と言う）の導入を公表したにもかかわらず、制度上の不備、識者・世論の猛反発により、2019 年 11 月 1 日、その延期が発表された。注意したいのは、民間試験の導入は中止ではなく延期であるという点だ。つまり、次回の新学習指導要領が適用される 2024 年度から民

間試験活用の実現が再度予定されているということだ。

そもそも、大学入試に民間試験を導入するという政策の起源はいつにあるのか。議論の出発点は、1984 年に中曽根内閣の下で設置された臨時教育審議会にある（江利川, 2013; 鳥飼, 2020）。1986 年 4 月 23 日に同議会で公表された「臨時教育審議会第二次答申」の中で、大学入試について、「英語の多様な力がそれぞれに正当に評価されるよう検討するとともに、第三者機関で行われる検定試験などの結果の利用も考慮する」という方針を示し、具体的な提案として「TOEFL などの第三者機関による検定試験の結果の利用も考慮する」と述べている。

このことからわかるように、大学入試における民間試験活用案は、今から 30 年以上前に提案されたものであり、長い歴史を有することがわかる。しかしその後、数年間にわたり民間試験活用については議題としては複数回あがったものの、それが実現されることはなかった。

民間試験活用について大きな動きがあったのは、第二次安倍内閣が発足した翌年の 2013 年であった。同年 4 月 18 日、自民党教育再生実行本部が「成長戦略に資するグローバル人材育成部会提言」の中で、「実用的な英語力を測る TOEFL 等の一定以上の成績を受験資格及び卒業要件とする」という提言がなされた。しかしこの提言では、TOEFL を「受験資格」として採用するのであって、現行のセンター試験英語を廃止し、「入学者選抜」として民間試験を活用するという案については一切議論されていない。

民間試験が「受験資格」から「入学者選抜」としての活用に移った発端は、南風原（2018）によると、2016 年 8 月 31 日、文科省が発表した「高大接続改革の進捗状況について」にあるとされる。同報告においてはじめて、「資格・検定試験の活用のみにより英語四技能を評価することを目指す」という方針が示された。そして、2017 年 7 月の公表につながりその実現化が目指されたものの、その 2 年後の 2019 年の 11 月に延期が発表された。

2.2. なぜ民間試験なのか

大学入試での民間試験の活用を目指す目的は何か。2016 年 8 月 31 日に文科省が公開した「高大接続改革の進捗状況について 1」では、その目的として、次の 3 点を挙げている。すなわち、①「グローバル化が急速に進展する中、外国語によるコミュニケーション能力（特にスピーキングとライティングの能力）の向上が課題」であること、

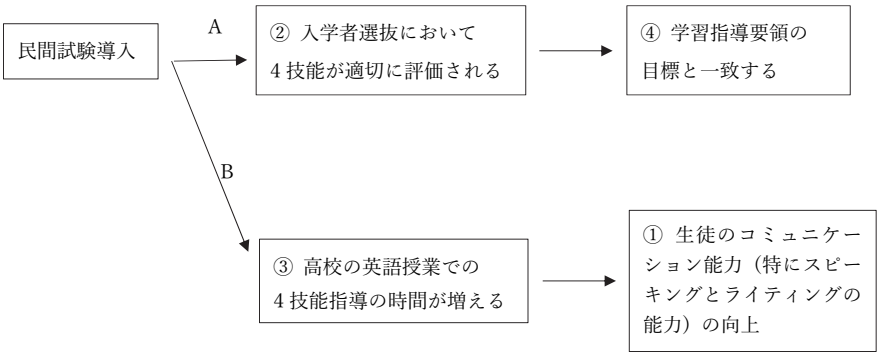
②「スピーキングとライティングを含む四技能評価の実現のためには、日程や体制等の観点から、民間の資格・検定試験を積極的に活用する必要」があること、③「4 技能評価の実現により、高等学校における授業改善を促進」すること、の 3 つだ(1)。

加えて、2017 年 5 月 16 日に公表された「高大接続改革の進捗状況について２」では、民間試験活用の目的について、「高等学校学習指導要領における英語教育の抜本改革を踏まえ、大学入学者選抜においても、『読む』『聞く』『話す』『書く』の 4 技能を適切に評価するため、共通テストの枠組みにおいて、現に民間事業者等により広く実施され、一定の評価が定着している資格・検定試験を活用する」と記されている。これは要するに、先述した②の目的を具体化させたものであろう。換言すれば、④「学習指導要領の目標との一致」(小泉, 2018) と定義できる。以上をまとめたものが図 1 である。

ここで注意したいのは、図 1 において矢印 A と矢印 B で表されているように、民間試験導入は性質の異なる 2 つの目的を有しているということだ。前者は、4 技能を重視する学習指導要領との整合性をはかることを目的とした「評価の観点からの推進論」なのに対し、後者は、大学入試に民間試験を導入することで、高校の英語指導の在り方や意識の変革をもたらし、生徒のコミュニケーション能力の向上を目指す「政策的な推進論」であるという点で、大きく異なる（寺沢, 2019）(2)。

図 1

民間試験導入の目的



本稿で注目したいのは、後者の「政策的な」観点から見た民間試験活用の効果である。つまり、入試を変えることでそれが教育関係者にどのような影響を与え、そしてそれが生徒のコミュニケーション能力の向上につながるのか、について検討することだ。

テストが教員の指導法や生徒の学習法に与える影響のことを、テスト論では「波及効果」（washback effects）と呼ぶ(3)。波及効果研究は、Alderson & Wall (1993) を嚆矢として様々な研究が世界各国で行われてきた。しかし、日本の高校英語教育を対象に、波及効果について検証した先行研究はきわめて少ない（Watanabe, 2013）。さらに、寺沢（2020b）で指摘されているように、波及効果に関する先行研究のほとんどが妥当性検証を目的としているため、それらの知見を政策的示唆につなげることはできない。したがって、現状、「入試を変えれば生徒の学力が向上する」ことを示すエビデンスは一切なく、関係者の主観的な経験に基づき議論が進められてきたと言っても過言ではない。そこで次節以降では、大学入試改革の「政策的な」観点から見た「波及効果」を検証するには、どのようなリサーチデザインが求められるかについて説明する。

3. 波及効果のリサーチデザイン

3.1. 波及効果を「政策的な」観点から検証するための方法論

政策決定に関する議論をする際に重要となるのが、その研究の「エビデンスの質」である。ここでいう「エビデンス」は、日常語の「根拠・証拠」とは大きく意味が異なる。寺沢（2020a） によると、「エビデンス」とは「特定の処置によって特定のアウトカム（結果、成果）が引き起こされると想定する因果モデル「処置→アウトカム」において、その因果関係を示唆する分析結果のことを指す」（p. 172）(4) 。

2 つの変数の間に因果関係が存在することを証明するには、「反事実」を考慮する必要がある。つまり、実際に原因が起こった「事実」と、原因が起こらなかった「反事実」を比較してはじめて、その真の効果がわかる。

しかし、因果推論の根本問題として、我々は両方の潜在的結果を観察することはできない（Holland, 1986）。つまり、実際に観察できるのは事実のみで、反事実については推測することしかできない。大学入試改革を例にすれば、ある人物が大学入試改革の前、または、後を同じタイミングで受験することは不可能であり、我々が観測できるのはその片方だけだ。そのため、実験デザインをどれだけ反事実に近いものに設定で

きるかがその研究のエビデンスの質を左右する。

こういったエビデンスの質に関する指標をまとめたものがエビデンス階層である。この概念は「エビデンスに基づく医療（EBM）」に由来するものだが、その基本的な考え方は教育政策にも充分応用できる。表 1 は、寺沢（2014）がオックスフォード大学・エビデンスベースト医療センターのガイドライン（Oxford Centre for Evidence-based Medicine, 2009）で示されているエビデンス階層を、教育政策の領域に翻案したものである。

表 1

教育政策におけるエビデンス階層

階層順位	エビデンス
1 （高い信頼度）	因果的研究（ランダム化比較実験）
2	コホート研究（追跡調査等）
3	介入を経験した人々を非経験者と比較した相関的研究
4	事例を集めたもの
5 （低い信頼度）	専門会の意見
	現場のデータに基づかない基礎科学的研究
	上記の調査のうちデザインが不適切なもの

（p. 20）

ここで示されているように、最も信頼性が高いのは、ランダム化比較実験（Randomized Controlled Trial: RCT）を用いて得られたエビデンスである。事実、教育政策の研究におけるランダム化比較実験（以下 ”RCT” と言う）の重要性は、ますます高まっている（Angrist & Pischke, 2009）。

しかし現実社会では、理想的な実験デザインを整えることは困難である。というのも、その実施には莫大なコストがかかるうえに、特に教育政策においては、倫理面での配慮に充分注意する必要があるからだ。

とはいえ、単に事例や専門家の意見を集約させたところで、その研究のエビデンスの質はかなり低いものであるため、政策的示唆につなげることはできない（寺沢, 2014）。

このような問題を解決するためには、代替的な方法が必要となる。その代表が、社会科学的な「比較」である。ここで意味する「比較」とは、RCT のそれとは違う。後者の場合は、比較をする前に無作為化等の手法を用いて理想的な条件を作るのに対し、前者の場合は、反事実に最も近い事例を選択し、それを観察することで、理想状態に近い条件設定を行おうとする（伊藤, 2011）。

津川（2016）は、観察データから実験に似た状況を作り出す準実験デザインの例として、以下の方法をあげている。

(1) 操作変数法（IV design: Instrumental variable design）
(2) 回帰分断デザイン（RDD: Regression discontinuity design）
(3) 中断時系列デザイン（ITS: Interrupted time-series analysis）
(4) 差の差分析（DiD: Difference-in-differences analysis, 差の差法ともいう）
(5) 傾向スコア・マッチング（PS: Propensity score matching）

（p. 49）

上記の準実験デザインのうち、本稿がテーマとする「大学入試の波及効果」の検証に有益なのは、中断時系列デザイン、差の差分析の 2 つであると考えられる。

まず、中断時系列デザイン（以下 ”ITS” と言う）とは、「時間を軸にして、あるタイミングよりも前か後ろかで介入の割り付けが変わることを利用する」（津川, 2016, p. 56）方法である。以下、英語入試改革を例に説明する。

2020 年度の英語入試改革の目的は、生徒のコミュニケーション能力（特にスピーキングとライティングの能力）を向上させることにあった。つまり、民間試験導入という介入によって、生徒のコミュニケーション能力に改善がみられる、という効果を期待していたことになる（図 2）。

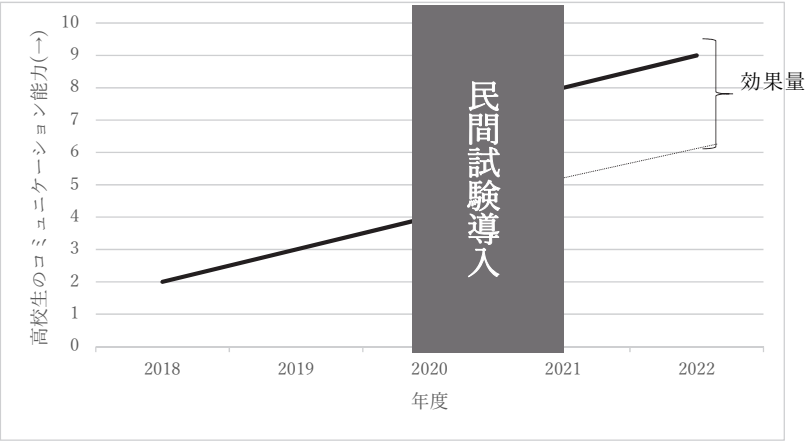
ITS では、介入前と介入後の変化に注目することで、その介入がもたらす効果を検証することができる。逆に言えば、介入前のデータなしに介入後の数値のみに注目するだけでは、変化量を知ることはできず、したがって、その介入の効果を知ることもできない。

大学入試改革の波及効果を検証する際にも、ITS に基づき介入後の数値だけでなく介入前の数値を計測し、その変化量を分析することが求められる。事実、波及効果に関する先行研究でも、ITS の考え方をベースに波及効果の検証をしているものは少なく

ない。しかし、その内のほとんどが後述する「リサーチデザインの条件」をクリアできていない。その点を解決しない限り、大学入試改革の波及効果を「政策的な」観点から検証することはできない。

図 2

英語教育改革推進派が想定する民間試験導入の効果



波及効果を検証する別の方法として、差の差分析（以下 “DiD” と言う）があげられる。DiD は、「介入群（たとえば、政策の影響を受けたグループ） と対照群 （政策の影響を受けなかったグループ）の 2 つのグループにおいて、政策導入前と導入後の 2 つのタイミングのデータを入手することからはじまる」（津川, 2016, p. 59）。

図 3 を使って説明していく。スピーキングテストが入試で必要な生徒を A（介入群）、必要でない生徒を B（対照群）とする。そして、A_1 と B_1 は、政策導入前のそれぞれのスピーキング能力、A_2 と B_2 は政策導入後のそれぞれのスピーキング能力を表している。

もし、「入試にスピーキング試験を導入することで、生徒のスピーキング力を向上させられる」という仮説を証明するには、図のように、A_2-A_1 の変化が、B_2-B_1 の変化より大きいことが示されなければならない。このように、DiD では、介入群と対照群の変化の差に注目することで、その政策のインパクトを解析する。

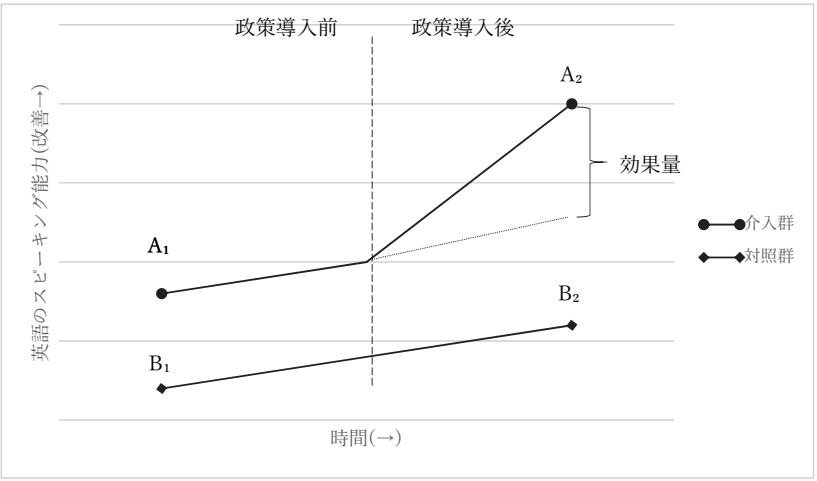
DiD は ITS とは異なり、同じタイミングで介入群と対照群の両方を必要とするため、

大学入試改革のような全国規模の制度的強制力を有するテストの波及効果を分析することには向かない。その代わりに、各大学で行われる試験や、地方ごとに行われる試験の波及効果の分析に関しては DiD の対象となる。

ただし、ITS と同様に DiD に関しても、リサーチデザインに工夫がなければ、たとえ介入群と対照群でその差分が有意なものであったとしても、エビデンスの質が高いとは言えない。また、政策実施前後の両方のデータがない場合、そもそもこの方法は使えないということも注意されたい。

図 3

DiD で政策効果が観察される場合



これまでの波及効果に関する先行研究でも、ITS や DiD の方法を採用しているものの、以下で述べるリサーチデザインの条件を満たせていないものがほとんどだ。次節では、ITS や DiD で得たデータを政策的示唆につなげるためにはどのようなリサーチデザインが必要か、過去の先行研究を参考にしつつ述べていく。

3.2. 政策的示唆につなげるためのリサーチデザインの条件

Watanabe (2004) では、波及効果を研究する上での方法論が記されている。その中で、

波及効果の分類として、以下の 5 つの指標があげられている（表 2）。

表 2

波及効果の分類法

1	Specificity	観察された波及効果はテスト全般に当てはまるものか、当該テストに限定されるものか
2	Intensity	その波及効果はどのくらいの強さを持つか
3	Length	その波及効果は短期的か長期的か
4	Intentionality	その波及効果は意図されたものか（意図されていない波及効果は発生していないか）
5	Value	その波及効果はポジティブなものか、ネガティブなものか

（pp. 20-21; 筆者訳）

上記の分類は波及効果の効果量について分析する際には重宝すると考えられる。しかし先述したように、リサーチデザインが適切に組まれていない限り、いくら効果量の大きい数値が観察されても、その知見を政策的示唆にはつなげることはできない。そのため、これらの 5 つの指標は、リサーチデザインの適切さが確認された後に、注目されるべき指標である。

寺沢（2018）は上記の 5 つの指標に加え、政策的観点から、さらに以下の 3 つの指標の追加を提案している。

1. テストは制度的強制力を持つか、非制度的なものか
2. 反実仮想モデルかどうか
3. validation のためのリサーチか、因果効果推定のためのリサーチか

大学入試を例に説明していく。大学入試共通テストは多くの高校生・浪人生が受験するハイステイクスなテストであるため、「制度的強制力を持つ」と言える。対して、例えば大学ごとに実施する個別のリスニング試験やスピーキング試験、あるいは、近年増加しつつある外部試験活用型の試験制度は、ある特定の受験生のみが対象となる試験であるため、「非制度的なもの」である。

また、前節で述べたように、2 つの変数の間に因果関係が存在することを証明するに

は、「反事実」を考慮する必要がある。しかし、大学入試改革のような制度的強制力を持つテストの場合、同時期に介入群と対照群を観測することは不可能なため、ITS や DiD といった準実験デザインを用いることで、反実仮想を考慮する必要がある。

3 つ目の指標については、「因果効果推定のためのリサーチ」でない限り、政策的示唆につなげることはできない。「因果効果推定のためのリサーチ」とは、「反実仮想モデル」をもとに仮説の検証を行っているか、という指標である。そのため、2 と 3 の指標は、ある程度重なる部分があると言ってよいだろう。

以上の議論を参考に、波及効果に関する研究を政策的示唆につなげるうえで考慮すべき条件をまとめたものが表 3 である。

表 3

波及効果に関する研究で考慮すべき条件

	考慮すべき条件	関連概念
1	因果効果推定のためのリサーチか	反実仮想モデル
2	内的妥当性が担保されているか	交絡因子の除去
3	外的妥当性が担保されているか	ランダムサンプリング

「因果効果推定のためのリサーチか」については、先述した通りの内容である。因果効果を推定するには、前提として反実仮想モデルが採用されている必要がある。ただし、現実社会で、介入群と対照群を無作為に割り当てることはなかなか困難が伴うため、準実験デザインを活用する必要がある。

また、エビデンスの質を高めるには、内的妥当性・外的妥当性の担保も不可欠である（寺沢, 2020a）。これらの説明については、中室・津川（2017）に詳しい。

> 内的妥当性とは、2 つの変数のあいだに因果関係のあることの確からしさを意味する。研究の対象となった集団に再度同じ介入を行った場合、同じ結果が再現される程度のことだ。一方、外的妥当性とは、研究の対象とは異なる集団に、その介入を行った場合、同じ結果が再現される程度のことを意味する。（p. 176）

内的妥当性を考慮するうえで重要となるのが交絡因子の存在だ。交絡因子とは、原因と結果の両方に影響を与える変数のことで、交絡因子が存在する相関のことを「見

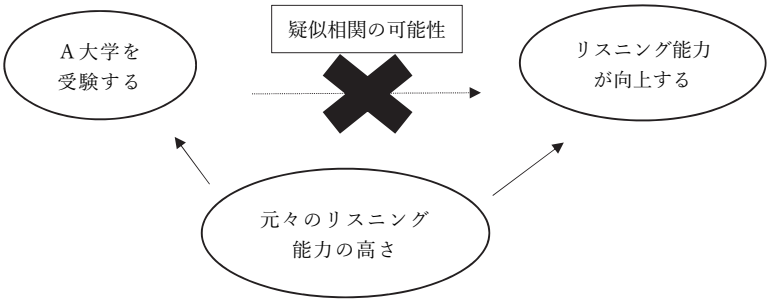
せかけの相関」「疑似相関」と呼ぶ。

波及効果に関する研究では、特に DiD を使用する際に、交絡因子の存在に充分注意を払う必要がある。例えば、試験内容として、リーディング・リスニング試験が必須のA大学とリーディング試験のみで受験可能なB大学があるとする。それぞれの大学の受験生を対象に調査を行い、その差異を分析したとして、その試験内容が生徒に与える波及効果（この場合はリスニング試験の有無）を正確にはかることはできるのか。仮にA大学の学生の方がリスニングの成績が高いとわかっても、その原因を「大学入試にリスニング試験があったから」と考えていいのか。

A大学を受験する高校生は、事前にその大学入試でリスニング試験が必須なことをわかったうえで、出願を決める。したがって、この時点で「ある程度リスニングが得意な高校生がA大学を受験しやすい」というセルフ・セレクションが発生してしまっている。つまり、「元々のリスニング能力の高さ」という交絡因子が発生しているため、疑似相関である可能性が高い（図 4）。

図 4

疑似相関の可能性

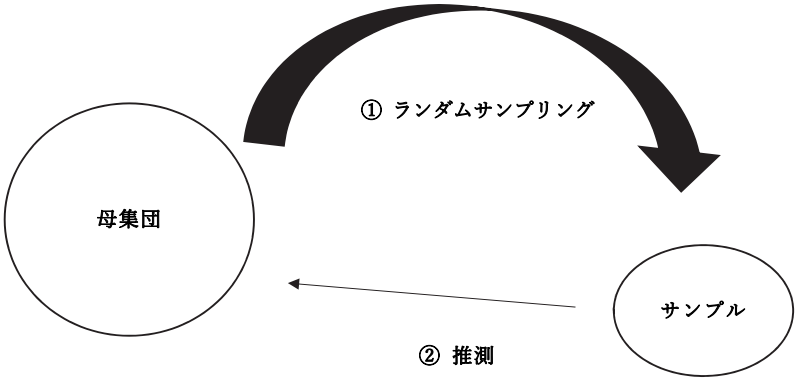


もちろん手元の限られたサンプルから母集団全体の真の値を正確に把握することはできない。しかし、そのサンプルが母集団から無作為に抽出されていれば、母集団全体の値を統計的に推測することはできる（図 5）。逆にそのサンプルが一部の学校を対象としている場合や、インターネット上の不特定多数の者を対象としている場合は、どんなに人数が多かろうとそのサンプルが母集団を代表しているとは言えず、したがってその結果を母集団全体に一般化することはできない(5)。

次節では上記にあげた指標をもとに、日本の高校生を対象とした波及効果に関する研究の問題点について述べる。

図 5

ランダムサンプリングによる推計統計



（竹内・水本, 2014, p. 46 を参考に作成）

3.3. 先行研究の問題点

本稿では、「2020 年度の大学入試改革、特に、民間試験の導入が高校での英語指導にどのような効果を与えるか」という波及効果を検証するために、どのようなリサーチデザインが適切かについてここまで述べてきたが、これと似た事例として、2006 年度の大学入試センター試験へのリスニング試験の導入があげられる。この入試改革の波及効果については複数の研究で分析がされている（Saida, 2009; Hirai, Fujita, Ito, & O'ki, 2013; 斉田, 2013）。

また、大学入試センター試験ほどハイステイクスなテストではないが、高校生を対象とした英語試験による波及効果を検証した最近の研究として、Sato（2018）があげられる。

これらの先行研究を、前節であげた指標をもとに整理したのが表 4 だ。それらの内、

ほとんどの先行研究が「反実仮想モデル」という点はクリアしている。また、ITS や DiD といった準実験デザインが使用されており、内的妥当性が担保されているものも多い。

しかし、「外的妥当性」に関してはどの研究も確保されていない。例えば、斉田（2013）では、関東地方にある一国立大学の入学者を対象に調査を行っており、この結果を日本人全体に一般化できるほどの外的妥当性を有してはいない。

また、Sato（2018）では、大学入試の際に、4 技能型の TEAP、2 技能型の TEAP、一般入試を使用した学生のそれぞれの英語学習への影響の差が検証されている。波及効果をどのようにして DiD で検証するかを理解するうえで有益な研究ではあるが、以下の理由からエビデンスの質は決して高いとはいえない。

まず、「4 技能型の TEAP を選択する受験生は、そもそも英語のリーディング・リスニングだけでなく、スピーキング・ライティングにそれなりの自信がある」という可能性が高い。したがって、試験でのスピーキング・ライティングの有無が英語学習に影響を与えた可能性以外に、もともとの英語能力が交絡因子として存在していた可能性がある。そのため、疑似相関の可能性があり、内的妥当性は低い。また先ほどと同様に、実験対象者が一部の私立大学に限定されているため、外的妥当性も低い。

表 4

先行研究の検討

	因果効果推定		内的妥当性	外的妥当性
	反実仮想	準実験デザイン		
斉田 (2009)	×	なし	×	×
Hirai et al. (2013)	○	ITS	○	×
斉田 (2013)	○	ITS	○	×
Sato (2018)	○	DiD	×	×

したがって、日本人の高校生を対象とした波及効果に関する先行研究で、上記の指標、特に外的妥当性が担保されているものは、筆者の知る限り存在しない。そのため、文科省が唱える「4 技能評価の実現により、高等学校における授業改善を促進する」という目標が果たしてどれほど現実的なものなのか、判断することは難しい。

この問題を解決する方法として、次の二つが考えられる。まず一つは、先述したラン

ダムサンプリングを行うことで外的妥当性を担保する方法だ。しかし、ランダムサンプリングは非常に強力な手法である反面、その実施にかかるコストも大きく、簡単に採用できる方法ではない。そこで、ランダムサンプリングを行うことが難しい場合のもう一つの方法として、事例研究の活用があげられる。もちろん事例研究では厳密な因果関係を推定することはできないが、理論的に意義のある事例を選択することで、質の高い研究となり得る（寺沢, 2019）。

そこで次節では、波及効果に関する事例研究を行う際に、考慮されるべき点について指摘する。

4. 事例研究の手法

事例研究とは、1 ないし少数の事例を詳しく観察する研究で、仮説の検証方法としても有効な手法である（伊藤, 2011）。しかし、野村（2018） が指摘するように、量的傾向の強いリサーチデザインを採用する際は、無作為割当や標本抽出に配慮して内的妥当性・外的妥当性の向上をはかるのに対し、「事例研究（特に質的なもの） については、どのように事例を選び、調査・分析していくかといった手順を論理的に考えて実施する人は少ない」（p. 42）。

そこで本章では、波及効果に関する事例研究をする際に、どのようなことに注意をすべきかについて、伊藤（2011）が提唱する方法論を参考にしながら、考察していく。

伊藤は事例研究を仮説の検証に使う方法として、9 つの使い道を提示している。その中から、波及効果の検証にも応用できるものとして、以下の 4 つがあげられる。

（1）あらかじめたてておいた仮説の検証
（2）その事例によって基本的原理を示す
（3）その事例によって因果関係に関する仮説や理論を作る
（4）変数が具体的にどのように関連するのかを突き止める

(pp. 119-122)

順に説明していく。(1) については、もっともイメージしやすい事例研究の使い道だと思われる。まずは理論に基づき仮説を立て、実際に観察された結果を比較することで、その理論の整合性を確かめる。ただし、この方法では検証のための事例が一つしかないため、比較や統計分析に比べると説得力が低いうえに（伊藤, 2011）、エビデンスの

質も低い。

（2）については、事例選択が重要なカギとなる。というのも、その事例が特殊なものではなく大多数を代表するものであれば、その事例を詳しく観察することで全体の傾向をつかむことができる。例えば、東京都では、都立高校受験に関して英語スピーキングテストの導入を 2022 年度から実施する予定を公表している。この事例は「東京都の中学生」に限った話であり、日本人全体に当てはまることではない。しかし、全国規模の調査と比べれば、はるかに調査を行うコストが低いうえに、東京都が他の都道府県と比べ、英語教育事情について「特殊」であるとは言えないだろう。よって、事例研究ではあるものの詳細に観察することで、「スピーキングテストを導入することで、学校の指導法・生徒の英語力に与える影響」の全体の傾向を把握することができると考えられる。

（3）のように、仮説を導く手段として事例研究を行うこともある。特に波及効果に関しては、日本の高校英語教育を対象に実施された先行研究がきわめて少ないため（Watanabe, 2013）、先行研究を概観してもどのような因果関係が存在するか不明な部分が多い。そこで事例を観察することで、具体的な因果関係のプロセスが明らかになり、仮説の構築が容易になる。

ここで注意すべきことは、ある事例を観察して導いた仮説を、同じ事例を用いて検証してはならないことだ（伊藤, 2011）。（1）で示した通り、事例研究は仮説の検証に用いられることもあるが、その際には仮説の構築に使用したものとは別の事例を用いる必要がある。

（4）に関しては、（3）の内容とやや重複するが、（3）が仮説の「構築」に焦点を当てていたのに対し、（4）はその仮説の「プロセス」に焦点を当てるという点で異なる。つまり、ある 2 つの変数の関係が確認された後、どのようにしてそれが生じているのか、あるいは、なぜそのようなことが生じたのか、という量的研究では把握しきれない側面の観察を目指す [6]。そのため、そもそも先行研究の数が少ない波及効果に関しては、（4）の方法を採用する前に、まずは量的研究や（1）の方法によって、仮説を検証することが求められる。

5. 結論

本稿は、2020 年度の大学入試改革における目標の一つとされている「4 技能評価の実現により、高等学校における授業改善を促進」、平たく言えば、「テストを変えることで高校の英語学習の在り方を変える」という目標を検証するために、どのようなリサーチデザインが求められるかについて論じた。

本稿で明らかとなったように、日本人の高校生を対象とした波及効果に関する先行研究は数が少ないうえに、そのリサーチデザインに問題があるため、良質なエビデンスが存在しない。そのため、この度の大学入試改革に賛成するにしても意義を唱えるにしても、根拠がほとんど無い状態で、いわば、各々の主観的な経験則に基づき、議論が進められているのが現状だ。

本稿は波及効果に関する研究の注意点を述べただけに過ぎず、重要なのは今後の研究でいかに高い質のエビデンスが得られるかである。具体的には、本稿で紹介した ITS や DiD 等の準実験デザインを用いることで、反実仮想を考慮した因果効果推定のためのリサーチが実現される必要がある。

もちろん、大学入試改革が高校生・受験関係者に与える影響は多岐にわたるため、「民間試験の導入」→「生徒のコミュニケーション能力が向上」という二つの要素だけでその本質を理解することはできない。その点で、インタビュー調査や事例研究といった質的な研究が重要であることは言うまでもない。

しかし、政策的示唆に直接つながるのは、やはり母集団を日本人全体とした、大規模な英語力調査であろう。特定の集団にアンケート調査やインタビュー調査を実施しても、エビデンスの質の高い研究とは決して言えない（寺沢, 2019）。

本稿から導き出される提言は次の二つである。まず、波及効果に関する調査結果を政策的示唆につなげる際には、少なくとも「反実仮想モデル」「内的妥当性」「外的妥当性」について十分考慮される必要がある。特に外的妥当性を考慮せずに、むやみにその結果を母集団全体に一般化するのではなく、そのサンプルはどの程度特殊なものなのか、あるいは、どれほど大多数を代表するといえるものなのかを詳述した上で、その結果が一般化できる範囲を適切に検討することが重要だ。

次に、入試改革前後の大規模な英語力調査の必要性だ。本稿で紹介した、ITS や DiD といった準実験デザインを使用するには、改革の前後のデータが不可欠だ。そのため、

長期的な視点を持ち、データ収集が行われることが求められる。2006 年度のセンター試験へのリスニング導入の波及効果を詳細に分析できなかった理由の一つは、そもそも大規模な調査があまり行われておらず、必要なデータが集まらなかったという原因もあるに違いない。このような事態を避けるために、今後の入試改革ではその前後のデータ収集をしっかり行い、数年後に質の高い研究が成されるための土台を固められるよう強く意識すべきだ。

付記

本稿を執筆するにあたり、貴重なご助言を下さった査読者・指導教官の先生方に心より感謝申し上げます。

注

(1) ① の「グローバル化」を教育改革の根拠とすることの問題点については、寺沢（2020a: 8 章） を参照されたい。

(2)「民間試験導入」→「② 入学者選抜において 4 技能が適切に評価される」→「④ 学習指導要領の目的と一致する」、といった「評価の観点」から民間試験導入を活用するには、テストの「妥当性」について考慮する必要がある。妥当性とは、「テスト開発者（テスト作成者）がテストで測りたいと思う能力（構成概念：construct） がどの程度測れているか、また、使用目的にどの程度合っているか」(小泉, 2018, p. 38) を示すものである。2020 年度の入試改革に関するテストの妥当性についての分析は、小泉（2018）, 宇佐美（2020）に詳しい。

(3) Cheng, Sun, & Ma（2015）によると、応用言語学では、テストが指導法や学習法に与える影響のことを「インパクト（impact)」または「ウォッシュバック（washback)」と表すことが一般的である。前者を「教室内での影響」、後者を「教室外での（社会的な）影響」と区別して使用する研究者もいる。

(4) この考え方は、医療分野に端を発する。エビデンスに基づく医療（EBM: evidence-based medicine）の考え方は 1990 年代以降、様々な分野に浸透した。「教育政策」もそこにルーツがある。この系譜については、寺沢（2014）に詳しい。

(5) 小林（2019）が指摘するように、RCT によって担保されるのは内的妥当性であって、外的妥当性は保証されない。そのため、RCT の結果をむやみに一般化してはならない。

(6) 野村（2018）は事例研究の重要性について次のように指摘している。「事例研究は、複雑な事象や新奇な事象を分析する上で力を発揮する。すなわちそれが「どのように」生起し展開しているのか、それが「なぜ」起こるのか、あるいはいったい「何が」起きているのか（存在しているのか） という問いに適したリサーチ・デザインである」(p. 46)。民間試験導入の波及効果を調べるには、寺沢（2019）が指摘するように、「高校生の四技能能力が向上したか」という指標のみを観察すれば、政策の効果自体は分析することができる。しかし、それに至るまでにどのようなプロセスが存在するのかは非常に複雑であるため、これを検証するには事例研究の蓄積が必要となる。

参考文献

阿部公彦（2018）『史上最悪の英語政策：ウソだらけの「4 技能」看板』東京：ひつじ書房

伊藤修一郎（2011）『政策リサーチ入門』東京大学出版会

宇佐美慧（2020）「記述式問題の現在：テスト理論から見た検討課題」中村高康（編）『大学入試がわかる本：改革を議論するための基礎知識』岩波書店

江利川春雄（2013）「『大学入試に TOEFL 等』という人災から子どもを守るために」大津由紀雄・江利川春雄・斎藤兆史・鳥飼玖美子『英語教育、迫りくる破綻』(pp. 1-27) ひつじ書房

江利川春雄（2018）『日本の外国語教育政策史』ひつじ書房

小泉利恵（2018）『英語 4 技能テストの選び方と使い方』アルク

小林庸平（2019）「エビデンスに基づく政策形成の考え方と本書のエッセンス」小林庸平（監訳・解説）『政策評価のための因果関係の見つけ方』日本評論社

斉田智里（2009）「大学入試センター試験リスニングテストの波及効果：茨城大学入学者の英語学習と大学英語カリキュラム改革の影響」『*JACET* 全国大会要項』, 48, 195-196.

斉田智里（2013）「大学入試センター試験リスニングテスト導入の高大英語教育における波及効果の解明（科学研究費助成事業研究成果報告書）」

竹内理・水本篤（2014）『外国語教育研究ハンドブック［改訂版］：研究手法のより良い理解のために』松柏社

津川友介（2016）「準実験のデザイン：観察データからいかに因果関係を導き出すか」『岩波データサイエンス Vol. 3』岩波書店

寺沢拓敬（2014）「英語教育学における科学的エビデンスとは？ 小学校英語政策を事例に」『外国語教育メディア学会中部支部外国語教育基礎研究会 2014 年度報告書』, 15-30.

寺沢拓敬（2018）「ウォッシュバック研究、7 月下旬に読んだ文献」https://terasawat.hatenablog.jp/entry/2018/07/31/153429 より所収（2020 年 11 月 26 日閲覧）

寺沢拓敬（2019）「『入試が変わらないから英語教育に成果が出ない』に根拠はない：政策効果の観点から見た『外部試験』議論」ひつじ書房ウェブマガジン『未草』http://www.hituzi.co.jp/hituzigusa/2019/02/28/letstalk-15/ より所収（2020 年 11 月 26 日閲覧）

寺沢拓敬（2020a）『小学校英語のジレンマ』岩波書店

寺沢拓敬（2020b）「『エビデンスに基づく教育』の可能性と限界」『現代思想』2020 年 9 月号, pp. 104-113

鳥飼玖美子（2020）『10 代と語る英語教育』筑摩書房

中室牧子・津川友介（2017）『「原因と結果」の経済学』ダイヤモンド社

野村康（2018）『社会科学の考え方』名古屋大学出版会

南風原朝和（2018）『検証 迷走する英語入試：スピーキング導入と民間委託』岩波書店

文部科学省（2016）「高大接続改革の進捗状況について１」https://warp.ndl.go.jp/info:ndljp/pid/11293659/www.mext.go.jp/b_menu/houdou/28/08/__icsFiles/afieldfile/2018/04/25/1376777_001.pdf より所収

文部科学省（2017a）「高大接続改革の進捗状況について２」https://warp.ndl.go.jp/info:ndljp/pid/11293659/www.mext.go.jp/b_menu/houdou/29/05/__icsFiles/afieldfile/2017/05/23/1385793_02_1.pdf より所収

文部科学省（2017b）「大学入試共通テスト実施方針」https://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2017/10/24/1397731_001.pdf より所収

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*(2), 115-129.

Angrist, J. D., & Pischke, J. S. (2009). *Mostly harmless econometrics.* Princeton University Press.

Cheng, L., Sun, Y., & Ma, J. (2015). Review of washback research literature within Kane's argument-based validation framework. *Language Teaching, 48*(4), 436-470.
Hirai, A., F, R., Ito, M., & O'ki T. (2013). Washback of the center listening test on learners' listening skills and attitudes. *ARELE, 24,* 31-45.
Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(3), 945-960.
Sato, T. (2018). The impact of the test of English for academic purposes (TEAP) on Japanese students' English learning. *JACET Journal*, *62*, 89-107.
Watanabe, Y. (2004). Methodology in washback studies. In L. Cheng., Y. Watanabe., & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods.* Lawrence Erlbaum Associates.
Watanabe, Y. (2013). The National Center Test for university admission. *Language Testing, 30*(4), 565-573.