

障害児教育における評価・アセスメントとその方法

— オンライン時代のための Differential Boost Hypothesis の再評価 —

貝原千馨枝
柴田邦臣

1. はじめに

ある教育がどの程度効果があるのか、その影響をどう評価するかについては、これまでも多くの研究者の関心を集めてきた。なかでも、多くのサポートや配慮が注ぎ込まれる障害児教育・インクルーシブ教育の分野においては、その教育内容の有効性はもちろん、コストとの比較衡量という観点から、さらには教育や支援の妥当性・公平性という観点から、常に注目を集め、そのための方法論の模索が続けられてきた。

その努力の一方で、教育評価の方法を定式化する試みは、未だ途上であるといえる。そもそも、評価方法の主要タームである test と assessment についても、十分な区別や定式化ができていない。例えば Kubiszyn & Borich (2015) は、test と assessment はしばしば置き換え可能な言葉だが、testing よりも assessment のほうが、評価する感が抑えられるために、アセスメントという言葉を使うこともあると主張している。Kubiszyn & Borich は、test と assessment は類義語として扱うものの、test と assessment と、assessment process は明確に区別する必要があると述べている。

本稿で言及するように、アメリカにおいても、1975年の全障害児教育法から障害のある子どものインクルーシブ教育が図られ、後述する IDEA や NCLB 法によってその加速度は増している。そのようにインクルージョンが進むのとは裏腹に、そのための学習評価概念やその手法については、十分な知見が集まっているわけではない。

本稿は、日本に数少ないインクルーシブ教育の学習評価論の経緯についてまとめ、その変遷から論点を整理し、障害児教育における評価の方法について考察する。その成果を受けて、今後の日本のインクルーシブ教育、特にCOVID-19 Crisis以降、急速にデジタル化・オンライン化がすすむ教育現場において、求められるアセスメントの理論的枠組みを提案する。

2. テストとアセスメントの変遷

アメリカにおける教育測定・教育評価の歴史は、これまでも時代を分けて語られてきた。例えば渋谷(2003)は教育測定・教育評価の変遷について、教育測定時代、教育評価時代、評価活動の現代化の3つの時代に分けて説明している。また北野(2015)も経済体制により、教育評価体制を3つの時代に区分している。しかしながら、障害のある子どもたちと教育評価との関係性の変遷はこれまでまとめられてこなかった。そこで特別支援教育や、インクルーシブ教育の観点から、これまでの教育評価の主要なトレンドの変遷を大きく3つに分けて示し、教育効果と合理的配慮の関係性について概観してみたい。

2-1. 第1期 measurementの開始：1900-1920年代

教育への評価・測定が開始されたのは、1900年代から1920年代である。この時代には、人間(子ども)が生得的に持っているといわれる知的能力を測るテストが次々に開発され、「学習者」の選別が行われていた。第1期の「学習者の選別」の象徴ともいえる動きが、1900年代からの「教育測定運動」と、1926年の「進学適性検査(Scholastic Aptitude Test)」開始である。

「教育測定運動」とは、20世紀初頭に見られた、それまでのテストへの反省から生まれた研究の動向である。教育測定運動を支えていたのは、19世紀後半に発展した自然科学の進歩を促した観察や、測定の手法を教育事象にも応用すべきだとする考え方であった。同時に教育の実証主義、教育統計の進歩、個人差心理学などにより、知能、学力などの客観的測定法が熱心に議論されるようになった(渋谷2003)。この教育測定運動には「標準テスト」「精神検査法」「科学的管理法」の3つの系譜があるといわれており(田中1988)、現在のSATの前身である1926年の「進学適性検査(Scholastic Aptitude Test: SAT)」開始は、この運動の集大成といえよう。そこで、まずは「教育測定運動」の3つの系譜を概観し、それらがいかに「進学適性検査」を実現させたの

か見ていく。

標準テスト

標準テストは、心理学者・教育学者の Thorndike, E. L. によって主張されたものである。Thorndike は、1904 年に『精神的・社会的測定学序説』を著し、「教育測定の父」と呼ばれている。「全て存在するものは、何らかの量において存在する」という言葉に象徴されている通り (Thorndike 1918)、彼はこれまでの論文体テストが主観的であり、結果に信頼が置けないことを証明するとともに、学力が客観的に測定できるものだと主張し、客観テスト形式の考案を促した (渋谷 2003)。当時の論文体テストにおける採点結果の研究は、1912 年の Starch, P. と Elliott, E. C. の研究が有名である。この研究では、142 人の教員に、中学 1 年の A, B 2 人の生徒の論文体の英語の答案を別々に採点してもらった。研究の結果、同じ答案に 60 点をつける教員もいれば、90 点をつける教員もあり、同じ答案でも教員によって採点が大きく異なることが明らかになった (渋谷 2003)。Starch と Elliott の研究にも見られるように、採点の客観性という問題への反省から、試験評価者の主観の入ったテストではなく、誰が採点しても同じ評価となる「客観性」を強く意識したテストが必要とされ、それが強くめざされる時期が続くことになる。

北野 (2015) は、この標準テストの系譜について、1911 年の「全米教育協会 (National Education Association: NEA)」の動向に注目しまとめている。NEA では標準テストの実施と点数の基準を策定する委員会「学校及び学校制度におけるテストと効率基準に関する委員会 (Committee on Tests and Standards of Efficiency in Schools and School Systems)」が設置され、1913 年から 3 年連続で検討課題の内容が報告されている。この委員会の目的は、学校の経営管理、児童の学力、教員活動の効率性などを測定調査する方法を検討し、学校の効率性を向上させようとするものであった。

この当時の標準テストは、質的側面を軽視するばかりか、その時の子どもの状態や、当時の教育課程、教科書などを肯定する保守的な理論により開発されたものであり (田中 1988)、またあくまでも学校の経営効率の向上のためのもので、個人の学力向上を目的として使われるものではなかった。

精神検査法

1905 年フランスの Binet, A. らにより、知能検査が開発された。Binet らが知能検査を開発した目的は、小学校に入学したのちに、学業についていくこ

とができない子どもたちをあらかじめ見分けることであった。Binetらの知能検査は、学び方を学ぶため、学ぶための態勢づくりのための検査であり、現在でいう特別支援教育のために知的検査を使おうとしていた。

Binetらによる知能検査は世界に影響を与え、当時隆盛していた優生学運動へ結びついていった。Binetらが自覚していた知能検査の限界や、あくまで実用のための検査であるという前提は無視され、アメリカではTermanがこの検査を参考に、1916年にスタンフォード-ビネー式知能検査を開発している。この知能検査は、1912年にドイツのStern, W.によって創案された知能指数 (IQ) 概念を実用的な検査に取り入れたものであった。このテストは全米的な規模で実施され、幼稚園児から第12学年までの子どもたちを対象に、知識 (academic knowledge) を評価することを試みた。Termanは、すべての人は客観的に予測可能な、知的能力があると主張した (Lemann 2004)。

「天才の発生的研究」においてTermanは、IQ140以上を天才とし、高いIQ指数の子どもたちの身体・精神・性格的特色をまとめた。Termanは「知能は1つの因子から成り立ち、遺伝によって生得的に決定されているため不変」であるとする立場をとっていた。IQの算出により、子どもたち個人個人にIQが与えられることとなり、子どもたちの様々な才能を、結果としてIQ概念のなかにのみ表してしまうこととなった。

さらに第一次世界大戦参戦により、Yerkes, R. M. が大規模な集団式の知能検査 (Army Test) を実現させ、検査結果を新兵の人種による比較分析に用いた。

人間の「知能 (intelligence)」を測定可能とした、1910年代半ばに登場する「IQ・知能テスト」は、たんに人間の知能を科学的に測定可能としただけでなく、人間 (子ども) の知能は生得的に決定されていて、人間の教育可能性の予測は早期に可能であるとしたものだった。さらには、生得的な知能水準と学業の間には深い相関関係があることも主張された (北野 2015)¹⁾。アメリカでは知能検査が優生思想と結びつき、激化していった。

科学的管理法

第3の系譜である「産業効率」をモデルとした「科学的管理法」については、「科学的管理法の父」と呼ばれるTaylor, F. W.が有名である。もともとは生産効率の向上と、工場における作業管理の合理化を目標としたものであったが、この科学的管理法の導入は、教育現場にも取り入れられるようになり、学校の経営管理に影響を与えた。

上述の通り、Yerkesによって開発された1917年のArmy Testでは「集団式知能検査法」が初めて採用された²⁾。知能検査の個別式から集団式への変化は、本来比較できないはずの人種などに焦点をあてた集団と集団の比較のための利用を可能にした。またこのArmy Testを契機に「多肢選択法」も開発された。これにより「採点の容易さ」「時間短縮」をもたらし、集団にも対応可能な、客観性を持って能力を測定するテストが実現・普及していった(北野2015)。

進学適性検査 (Scholastic Aptitude test: SAT)

Termanを中心とする心理学者による影響は、学校教育にも強くもたらされたという(Ravitch 2000=2008)。Termanは、精神年齢を用いることで、児童・生徒を均質の集団に割り振り教えることが可能だと説明し、就学者の急増に伴う学校運営上の問題を解決する手段として、集団知能検査を提唱した(Terman 1919, Terman 1923)。

1926年には、現在にも行われている大学進学のためのテストSATの前身「進学適性検査 (Scholastic Aptitude Test: SAT)」が開発された。SATとは、アメリカの大学に入学する際に考慮する要素の1つである。このSATの開発当初には、Termanの主張を踏襲するBrigham, C. C.やYerkesが関わっており、「学びに対する適性を確かめることを目的とする知能検査」(Ravitch 2000=2008)となっていた。開発者の1人である、心理学者(優生学者)Brighamは、高等学校の科目の達成度よりも、知的能力に基づく受験者の特定のほうが、その後の高等教育の成功を予測できると主張した(Crouse 1988)。このテストは、判断、計算、場合分け、人工言語、反対語(対義語)、数列、論理的推論、そしてリーディングの問題で構成されていた。この当時のSATは、315の設問に、90分で回答するものであり、300以上の試験会場から800人以上の受験者が受験したという。SAT同様、大学入学時のテストとして、当時College Board³⁾が「共通試験」を開発しており、これは「これまでの個別性の高い入学試験と比較して格段に優れた選抜試験」だったというが、「共通試験」の受験には長期的な準備が必要であった一方で、SATは全国各地に試験会場を設置し実施するため、地方の高校からでも受験が可能であり、様々な地域や階層から大学入学者を選抜することができるという利点があった(荒井2003)。第二次大戦後、G・I法⁴⁾の影響を受けてCollege Boardへの加盟校が大幅に増え、SATの受験者が爆発的に増加した。1945年に2万6000人であったSAT受験者数は、1955年には8万6000人に、1965年には10倍の85

万人に、そして1970年には100万人を越えたという (Schudson 1972)。SATの“Aptitude (適性、才能、素質)”という言葉が象徴しているように、この時代の教育評価は、努力や工夫ではなく、生得的な知能や適性を測り、それによって学習者や進学者を選別していた。教育測定運動によるSATの誕生は、当時の大学就学者数の急増にいかに対応するかという問題を背景としており、知能検査を行うことで「(学びに) 向いている者」「向いていない者」をあらかじめ把握するものであった。またそれが集団式の知能検査であったことから、本来比較できないはずの人種集団を比較するなど、生得的な要因による「学習者」の選別を可能にする試験だったといえる。この時期に開発されたテストは、どのように教えたのかといった学習や教育の具体的な内容や生徒の習得度というよりも、実施された教育そのものに、どの程度の影響があったのかを測定できるとし、数値的に把握することが可能になったという点に、まさに特徴があるといえることができる。

その後、1930年代にはSATにおいて障害のある受験者への配慮が実施されるようになった (ETS 1988)。1937年には、SAT (Scholastic Aptitude Test) において視覚障害の受験者への配慮が提供され、College BoardはAmerican Foundation for the Blindの支援により、100の反対語、50の類推問題、そして50の読解問題が入った点字のブックレットを開発している (Rosenfeld et al. 1995)。1956年には、U.S. Civil Serviceによる、昇進に係するテストにおける、視覚障害の受験者への配慮の、テストの妥当性への影響が調査され、また同時に視覚障害の受験者への試験が実施された (Pitoniak & Royer 2001)。

また、配慮の基準をリスト化した、1985年版The Standardsの前身となるTechnical Recommendations for Psychological Tests and Diagnostic Techniques (APA 1954) が発表され、第二版 (Standards for Educational and Psychological Tests and Manuals) は1966年、第三版は1974年に出版されている。

2-2. 第2期 evaluationの導入：1930年代-1975年頃

Scholastic Aptitude Test: SATに象徴されるように、教育測定運動によって誕生したテストは、生得的な要因から「学習者」を選別するものであった。その教育測定運動によって生まれたテストが、単に客観的に現状を測定し、数値化するのみにとどまっていたことへの反省から、教育測定ではなく教育評価へと教育観が大きく変化していった。教育評価概念は1933年から行われた8年研究により誕生した概念である。この8年研究は、環境的な要因により、生徒・学生の学力が変化する可能性を示すものであった。さらに

1930年代には上述の通り、SATで視覚障害の受験者に対する配慮が行われ、点字を使用しての受験が可能になった。1930年代から学習に関して、環境的な要因が注目されるようになっていく。

8年研究(measurementからevaluationへ)

教育測定運動は、客観的に個人を測定する「教育測定(Educational Measurement)」が「もっと教育の目標と緊密に結びつけるべきなのにその自覚が希薄だった」こと、そして「知識・技能といった測定・量化しやすいものを重視し、教育の成果や教育のプロセスの評価が脱落していた」ことに対する反省により、新たに「教育評価(Educational Evaluation)」運動が展開されていった(渋谷2003)。教育評価運動の直接的な契機となったのは、アメリカの進歩主義教育協会(the Progressive Education Association: PEA)が実施した「8年間研究(The Eight-Year Study)」である。この研究は教育学者Dewey, J.や、後述する全国学力調査に議長として参画することになるTyler, R. W.らが主唱者として結成された人々により推進された。彼らは、アメリカの伝統的な入学試験制にもとづく入学者グループと、入学試験なしで入学させるグループをつくり、高校から大学へ入学したのちにどうなるかについての追跡的研究から、高校のカリキュラムについての実験を行なった(渋谷2003)。この研究の背景には、当時のアメリカの中等教育カリキュラムが、大学入学のニーズを満たすように設計されているという批判があった。この研究は、大学入学のためのテストを受験するという入学条件を失くし、高校の校長の推薦で大学へ入学させることを認めさせた(荒井2003)。当時、1929年に起こった大恐慌により大学入学者が減少していたため、大学側は比較的容易にその要件を受け入れ、テスト受験という条件がなくなったことで、高校側はカリキュラムの実験を開始した。入学試験なしで大学に入学したグループの高校では、より学際的な内容で、かつ芸術や課外活動に重点を置いたカリキュラムを学び、大学へ入学した。その結果、入学試験を経験したグループと同じ成績をとったが、それに加えてさらに芸術、政治、社会的な活動を積極的に行ったという結果が見られた。設備の整った、社会階層の高い学校の生徒は、カリキュラムリソース(コンサルタント、助成金、宣伝)によってより優れた成績を取めたことから、カリキュラム構造が、よりよい条件下(家族や教員がリベラルなイデオロギーを持ったとき、大学が審査を緩和したとき、財団が支援を提供したとき)で変化する可能性があることをこの研究は示した。第二次世界大戦と冷戦の影響によって、研究は8年で終了したが、この改革は

大きな影響を与え、課外活動等による教員の労働力の増加や、大学の競争が激化した(Tyack 1995)。

環境要因により生徒・学生の学力が変化するという8年研究から、学習観が大きく変化した。さらに、1969年から現在に至るまで実施されている全国学力調査(National Assessment of Educational Progress: NAEP)に、教育測定から教育評価への転換を見ることができる。

全国学力調査(National Assessment of Educational Progress: NAEP)

1965年には、初等中等教育法(Elementary and Secondary Education Act: ESEA)が成立した。これは経済的・社会的に不利な状況にある児童・生徒が、一定のレベルの学業を達成するため、州に与えられる連邦補助金制度を定めたものである。また同年、全米学力調査(National Assessment of Educational Progress: NAEP)が創設された(Fitzharris 1993)。NAEPとは、アメリカで定期的に行われる全国学力測定である。1969年から1979年まで毎年実施され、1980年から96年の間は2年に1度実施され、再び97年からは毎年行われている。NAEPには2つの目的がある。1つめのMain NAEP(主調査)は、最新の教育測定技術の適用と開発を行う調査で、社会的要請の高い教育課題へ取り組むものであり、もう1つのLong-term Trend NAEP(動向調査)は、時系列的な学力変化を調べるものである。被験者は4学年、8学年、12学年となっており、2000年のMain NAEP(主調査)では2500校から10万6500人の被験者が抽出された(村木 2003)。試験科目は読解、数学、理科、作文、米国史、公民、地理、芸術など広範囲に渡っており、音楽、演劇、美術、舞踊などを含んだ芸術教科も最近では測定している(荒井 2008)。被験者をNAEP側がサンプリングしている通り、NAEPは個人の学力には関心がなく、アメリカの地理的、あるいは民族的な様々な集団のその時点での学力分布とその変遷を調査することを目的としている。そのため生徒個人や学校単位のデータにはアクセスすることはできないが、これによって調査結果が特定個人の学力評価や、教員の能力評価に短絡的に使用されることを避けている。しかし、調査後の報告書は公開され、例えば性別、人種(民族)別、保護者の学歴別統計や、保護者、教師、委員、関係者などへのQ&A、そして教育や授業に役立つ教員・教材開発者向けの資料、児童・生徒、教員、校長などから得た背景情報と、成績との関連分析情報の要約などがまとめられている。

「アメリカの子どもたちの全体の達成度」を測るNAEPは、受験者の学力のみを測ってその現状を把握することで、純粋な教育効果を測ることができ

る。また、NAEPはあくまで達成度を測ることを目的とし、そこでの個人の成績は公表されないため、大学入学のための選抜に用いられるSATのようなテストとも性格を異にしている。第2期に始まったNAEPは教育測定から、教育評価 (evaluation) へと、教育の効果を把握する視点の変化を象徴するものの1つであるといえる。

NAEPはその後、1990年の個人教育法 (Individuals with Disabilities Education Act) の影響を受け、障害のある子どもたちを含めた調査を実施している (Beaton et al. 2011)。

障害のある子どもの教育参加

1940年代には、特殊学級・特殊学校に在籍している子どもの数は、障害のある子ども全体の1.56%であった。しかし、人種に基づく分離は差別であるとした1954年のブラウン判決が、1970年代の障害のある子どもの教育へと大きく影響していく。ペンシルバニア州での、1971年PARC裁判⁵⁾と1972年ミルズ裁判⁶⁾では、それぞれ知的障害のある子どもが教育プログラムから利益を得られること、重度の障害のある子どもに対する教育の必要性が確認された。その結果、1970年に特殊教育規定を備えていた州はたったの13州に過ぎなかったが、1974年までに35州へと増加した。しかし、実際は1960-70年代初頭になっても全ての子どもの就学義務が達成されたわけではなく、障害のある子どもは公教育から排除されていた (吉良 2015)。

2-3. 第3期 assessment への注目：1975年頃から現在

Scholastic Aptitude Test から Scholastic Assessment Test へ

1970年代の教育における evaluation の導入は、広範囲な影響力をもち、現在でも教育の改善に一定の役割を演じている。しかしこれ以降、80年代に入ると、教育における能力評価は、2つの点で変容を迫られるようになる。

第一に、「assessment化」とでも呼ぶべき変容である。1983年の“A Nation at Risk” (『危機に立つ国家』⁷⁾) により、アメリカの経済的な衰退と学力の低下が公に発表されたが、1960年代からSATの得点が降下し、70年代に入っても学力低下の原因が解明できないままであった。このような状況を受け1990年、SATは改革を実施した。まず、1941年以来変わっていなかった基準点を再調整し、出題内容や解答方法にも大幅な手直しを加え、またこれまで時間制限の厳しい試験であったが、30分延長することが決定された (荒井 2003)。最も大きな変更は、これまで進学適性試験 SAT (Scholastic Aptitude

Test)として実施されてきたテストを、進学能力評価試験 SAT (Scholastic Assessment Test)と改称したことであった。しかしながら、SAT から“Aptitude”という語を消し去り、生得的な能力を測定するという意味から“Assessment”へと改名したものの、その内実はいまだに生得的な能力を測る知能検査を含むものであった。

2000年、カリフォルニア大学の入試改革を指揮してきた総長 Atkinson, R. は、SAT⁸⁾がいまだに知的検査の色合いを残していると訴え、カリフォルニア大学の入学要件から外すように関係者へ指示した。SATの年間受験者は、延べ220万人で、そのうちの13%がカリフォルニア州の受験者が占めていたため、SATを主宰していた College Board は Atkinson の主張を退けるわけにはいかず、2002年にSATの大幅な見直しが発表された(田中 2002)。この見直しにより、SATの言語テスト、数学テストの両方から知的検査的項目を外し、言語テストには長文読解と文法の問題が加えられ、数学のテストではより難易度の高い問題を出すことが決定された(荒井 2003)。これらの変化は、アメリカの教育上の能力評価が、能力を競わせるテストではなく、一人ひとりの達成度を見る、アセスメントとしての役割を帯びてきたことの証左であろう。

教育関連の法律の成立・改正と IEP

もう一つの変化は、障害のある子どもたち向けの支援のあり方である。1970年代によく確立してきた特別支援教育は、ここにきて一つのハイライトを迎える。それが、1975年に施行された全障害児教育法 (Education for All Handicapped Children: EAHC)であった。EAHCは「障害児の権利章典」ともいわれる法律であり、障害のある子どもたちの教育に極めて大きな変化をもたらすものであった。この法律により、これまで「分離すれども平等」とされてきた障害のある子どもの教育について、全米で800万人の障害のある子どもたちへの公教育を受ける権利が保障された。この法律の主要な点は、3歳から21歳までの子どもが、無料かつ適切な公教育を受ける権利を有しているとされたことと、障害のある子どもが最も制約の少ない環境で教育を受けることを公立の学校に義務づけたこと、そして個別教育計画 (Individual Education Plan: IEP) が導入されたことである。

IEPは、障害のある子どもの保護者を含めたチームで、個別化された教育計画を作成する。この教育計画のなかには、子どものその時点での教科ごとの達成度や、機能的な遂行能力、年間達成目標などが明記され、都度見返す

ことが義務とされ、法的拘束力を持っている⁹⁾。IEP 導入は、これまで焦点が当てられてこなかった、個人の学習の向上を目的としているところに大きな違いがある。

その IEP の導入が、障害のある子どもたちにとって重要であったのは、それが子どもたち一人ひとりに対して計画されるものであり、子どもたちどうしを比較したり、データを集めて分析するという仕組みとは異なった、新しい性格を備えていた点である。その意味で、measurement から evaluation へという教育効果の変遷とは別に、assessment という新しい教育評価の方向性が生まれていることを、確認することができる。

しかし一方では、1994 年の初等中等教育法により、アメリカ国民全体が読み書きできるようになることが強調され、全州学力試験の実施と、その報告が義務づけられたり、学校単位で成果に対する報酬と制裁が行われることとなった。この法律の改正によって、障害のある子どもが学力試験を受験した場合、学校の総得点が下がるために、州によってはテストから排除されることもあったという(後上, 中澤, 松村 2002)。

このような状況のなか、1990 年に全障害児教育法は IDEA (Individuals with Disabilities Education Act) へと改称・改正された。IDEA では、障害のある子どもが通常のカリキュラムから排除されてきたこと、それによって、障害のある子どもが高い学力を達成することが制限されていること、また障害のある子どものドロップアウトが、障害のない子の倍みられていること、そしてその多くの就職が困難であるなど、いくつかの教育に関する課題があげられていた。そこで、この法改正では、障害の有無にかかわらず、すべての児童・生徒への同じカリキュラムへのアクセスの提供が強調された。1975 年の全障害児教育法から IDEA への変化は、その名前の通り、個別の障害者の教育法になったことであった。この法律の整備によって、一般学級において、インクルーシブな環境で学ぶ子どもたちが増えた。これらの動きは、生得的な要因で学習者を選別する第 1 期の measurement や、環境要因との関係が注目されるようになった第 2 期の evaluation 概念とはまた異なった、新しい評価概念である assessment の成立という動きと軌を一にする。

assessment において重要なのは、一般学級において、その子ども 1 人に対していかに配慮を提供し、障害のない同級生と一緒に学ぶかということである。一方で、evaluation というべき評価の仕組みとの併存も、第 3 期の特徴でもある。2002 年には、NCLB 法 (No Child Left Behind Act) が成立・実施された。この NCLB 法は、学力格差を是正するために制定された初等中等教

育法の改定法であり、特に低所得家庭の子どもたちの教育を向上させることをめざしたものである (U.S. Department of Education 2002)。この法律により、全州において全ての子どもたちが2014年までに読み書き・算数(数学)の能力が「習熟レベル (proficient)」に達することを求め、かつ州テストの実施が義務化された。学区・学校に対しては、テスト結果のアカウントビリティ(説明責任)を課し、目標基準値を達成できなかった学校や教員へ制裁措置 (sanction) として改善を求めた。「要改善」が2年以上続くと「罰則規定」が適用される。この規定の内容は、その学校の生徒への「学校選択権」「補習教育サービスの提供」「学校改善のための是正措置」などである。5年以上「要改善」が続くと、学校再編のための「Charter School への移行」「教職員の入れ替え」「管理運営を州政府や民間企業に委託」といった措置が義務付けられている。つまり、テストによって生徒個人個人の学力がどうすれば向上するかということよりも、テストによって成績の悪い学区・学校を明らかにし、主に経営面での改善を求めることが、NCLB 法で義務化されたテストの特徴であり(北野 2017)、その意味で、measurement ではなく evaluation という位置付けが定まったといえる。

この NCLB 法成立から2年後の2004年、再び IDEA が改正された。2004年の IDEA は、障害のある生徒の教育への期待度が低いこと、そして教育法・学習法に関してエビデンスに基づいた研究が十分になされてこなかったことを明らかにし、NCLB 法と密接に対応させ、障害のある児童・生徒が高い期待度をもって受け入れられ、可能な限り一般の学級で通常のカリキュラムに則った教育をめざすものとされた。このカリキュラムによって、2003年以降、障害のある児童・生徒(6-21歳)の半数以上が、80%以上の時間を一般学級で過ごしている (U.S. Department of Education 2019)。

第3期の法律の成立や改正は、他者や他地域、他の学校と学力を比較するための evaluation を続けつつも、一方ではこれまで学びやテストの受験から排除されてきた障害のある子どもたちの学習環境を大きく変えるものであった。IEPの誕生により、個別の子どもたちの学びをいかに深めることができるかに焦点が置かれ始めた傾向は、成績や効果を比較するテストではなく、その子ども一人ひとりの成長や達成度にこだわって時系列に評価するという、新しい評価の手法の提案につながっている。これが assessment という呼称の定着と軌を一にしている。assessment はテストの結果(得点)を評価するというだけでなく、あくまで個人がどのくらい伸びたり、どのような配慮を実施するとより制約を受けずに学びを深められるかを測定するものな

のである。

これまで、Kormos (2017) などにもみられるように、テストと assessment は、概念的には明確に区別されてきたわけではなかった。しかしここで私たちは、measurement から evaluation という流れで発展してきたテストと、IEP から新しく生まれた教育評価方法である assessment を区別して議論することができるようになるのである。

表 1：学習への評価・アセスメントの歴史

特徴	年	出来事
第 1 期 measurement の開始	1905	仏・Binet ら 知能検査開発
	1911	全米教育協会 NEA による標準テスト実施の検討開始
	1912	米・Terman 知能検査開発
	1912	Starch と Elliott の研究 (採点の客観性)
	1917	米軍隊で集団知能検査実施
	1926	進学適性試験 SAT (Scholastic Aptitude Test) 開始
第 2 期 evaluation の導入	1933-41	8 年研究
	1960 年代 (-90 年代)	SAT の得点が降下
	1965	全国学力調査 (NAEP) 開始
第 3 期 assessment への注目	1975	全障害児教育法 EAHC 施行、個別教育計画 IEP 導入
	1983	危機に立つ国家 “A Nation at Risk”
	1985	The Standards for Educational and Psychological Testing 改訂
	1994	初等中等教育法改正
	1997	EAHC から障害のある個人教育法 IDEA へと改称・改正
	1990	進学能力評価試験 SAT (Scholastic Assessment Test) に改称
	2000	カリフォルニア大学の入学要件から SAT (知能検査を含む) が外れる
2002	No Child Left Behind 法成立	

3. 教育効果と配慮効果

3-1. 学習評価

ここまで、子どもたちの学びの評価をめぐる象徴的な概念である measurement, evaluation, assessment という言葉を用いて整理してきた。現在では、子どもたち1人1人に対する教育や学習に関する assessment が実施されている。しかし、障害のある子どもたちについて考えたとき、具体的にどのように assessment を行うかについてはまだ結論がついていない。なぜなら障害のある子どもが一般学級の子どもと一緒に学ぶ時、何が配慮で、何がそうではないかを判断することは、容易ではないからである。そこで第3節では、これまでの、学習への妥当性と公平性への、配慮の影響に関する議論と、障害のある子どもへの配慮が、いかに学習の妥当性や、公平性に影響を与えているかをまとめた3つの仮説を紹介する。

3-2. 配慮と妥当性・公平性

学習において、配慮を提供した際の妥当性と公平性への影響が最も問題になるのは、test の場面である。test における妥当性と公平性は、インクルーシブな学習環境が整備されはじめた20世紀末頃から盛んに議論されるようになった。例えば、1985年には the Standards for Educational and Psychological Testing (AERA/APA/NCME 1985) が出版され、障害のあるテスト受験者への配慮に関する8の基準がリスト化され、現在も改訂が続けられている。assessment においてもっとも重要なのは、テストが測りたいものを正確に測れているかというテストの妥当性である。Henning (1987) は、妥当性を次のように定義している(訳は筆者によるものである)。

妥当性とは、与えられたテストや、そのテストを構成する部分が、測定したいものを測定する方法として適切かどうかを意味している。いずれのテストも、意図して測定しようとしているものを測定する範囲内では妥当であるといえる。“valid”という用語が、あるテストが妥当だと示す目的で使われるとき、基本的にはそこに前置詞の“for”が伴われるべきである。どのようなテストでも、そのように前置詞が伴われれば、他の目的にとってはそうではないにしても、ある目的にとっては妥当であるといえる。(89)

Henning の議論を引き継ぎ、Messick は妥当性を脅かすものとして2点を挙げている。1つは、テスト自体の構造化が不十分で、そのテストで主として測りたいものと関連している部分に焦点が当たっておらず、測りたいものを測ることができていない場合である(例えば、ライティングの力を測るためのテストで、ディクテーションをしてしまうなど)。もう1つは、テストの構造には関係のないものが含まれており、中心とするべき要素ではない要素ではない要素も含めて測定してしまっている場合である(例えば、ライティングのテストの評価基準に、文字を秩序立って手書きすることができるといふものも含まれているなど)。近年 test の分野では、妥当性と公平性の関係が注目を集めている(Davies 2010; Kane 2004, 2010; Kunnan 2010; Xi 2010 など)。例えば公平性は妥当性から完全に独立しており、関係がないとするもの(Joint Committee on Testing Practices 1998, 2004; ETS 2002)もあれば、対照的に公平性と妥当性は両立しがたい構成要素だとするもの(Davies 2020)、また公平性を、妥当性を含む包括的な概念であるとするもの(Kunnan 2004)、そして the Standards for Educational and Psychological Testing (AERA/APA/NCME 1999) は公平性を妥当性と距離があるものとしつつ、強い相関関係があるとするものなどがある。

議論の焦点は、(合理的な)配慮を提供する場合にその配慮が、学習者やテスト受験者のうち、特定のグループに対してそこまで公平な学習・受験を保証しうるのか、その配慮がどのように評価の妥当性に影響するのかということである。テストの内容や、テストの実施手順において、配慮や変更をどこまで承認するかは、テストや学習内容の妥当性に影響を与える可能性がある。

しかしながら、何が配慮(accommodation)で何が変更(modification)なのかの区別は、困難な場合がある。例えば言語テストの分野においては、Hansenら(2005)が、あるテストの課題の種類と、それが測定しようとしている構造との関係性について、どう配慮するかを調査に関する分析枠組みを作成している。Hansenらはこのモデルを実証するために、読解力を測る課題で読み上げの配慮を受けている、ディスレクシアの受験者の事例を用いた。リーディングテストを構成しているのは、デコーディング(活字を音声化できるか)と、理解(音声化したものを理解できるかどうか)の主に2点であり、ディスレクシアの受験者は、書かれているテキストをデコーディングする必要がないために、読み上げの配慮は不当にディスレクシアの受験者にアドバンテージを与えてしまう。このように、読み上げはテストの構成概念の妥当性に影響するため、Hansenらはこれは配慮ではなく、変更であると主張して

いる。一方で、読解に関する問題に筆記ではなく口頭で解答することは、テストで測りたいリーディング力の要素の構成に影響せず、またこの調整であれば試験の妥当性に影響しない。Hansen ら (2005) のモデルは、high-stakes の試験でのアセスメントと、教室でのアセスメントの両方に適用することが可能であり、言語の試験における配慮と変更を区別することができ便利である。また、Phillips (1994) は、受験者に対し配慮を提供すべきかどうかを判断する際の検討事項をリスト化している。そのうちの一部が、配慮と変更の問題に関係している。

- その配慮は、テストが測定したいものの要素の構成を変えてしまうか？
- その配慮は、テストの得点の意味を変えてしまうか？
- その配慮は、障害のない受験者もなにかメリットを得るか？

この問いに答えるためには、配慮を実施した場合とそうでない場合で、障害のある学習者と、そうではない学習者のパフォーマンスを比較する必要がある。これまで配慮を提供した場合の、障害のある受験者とそうではない受験者の得点率については、様々な調査が行われ、その結果がまとめられてきた。障害のある子どもと、そうでない子どもとの間の得点率や回答数の増加を比較し、配慮の妥当性と公平性について検討した主要な3つの仮説が、Maximum Potential Thesis, Differential Boost Hypothesis, そして Interaction Hypothesis である。

3-3. 配慮に関する仮説

まず1つめの仮説は、Maximum Potential Thesis (Runyan 1991; Zuriff 2000; Lovett 2010; Stretch & Osborne 2005) である。この仮説は、後に説明する Differential Boost Hypothesis と、Interaction Hypothesis よりも前に発表された仮説で (Jennings 2014)、次のような例で示される。

学習障害のある子どもとない子どもの、試験時間延長の配慮を実施した際の得点と、そうでない時の得点の比較を行った。その結果、学習障害のある子どもは試験時間を延長した結果得点が伸びたが、一方で学習障害のない子どもには得点の向上は見られなかった (Zuriff 2000)。

このように、特定の介入の結果、障害のある子どもの得点は向上するが、

障害のない子どもの得点には影響がないという場合が Maximum Potential Thesis に該当する。Zuriff (2000) は、この仮説の有効性を調査するため5つの事例を調査した。しかし、全ての事例がこの仮説に該当したというわけではなかったため、仮説を一貫して支持できるエビデンスは見つかっていないと結論づけている。

障害のある子どもにとってメリットがあるが、そうではない子どもにとってはそうではないという状態は、つまりその介入が純粋に配慮効果のみを与える場合ということになり、その介入による「そのほかの効果」(仮に余剰効果と呼ぶ)は0であるということの意味する。障害のある子どもとない子どもがともに学ぶ場面を考えたとき、この仮説の非現実性が際立ってくる。例えば、授業において純粋に障害のある子どもへの配慮効果のみがあらわれるのは、60分間の授業で、30分ぶんの内容しか扱わなかった場合ということになる。この仮説は、純粋な配慮効果のみを測定することで、その差分が厳密に配慮であることが証明できる。しかし、この想定は、Kormos (2017) らが論じるように、やや非現実的な想定であるといえる。どのような配慮であっても、全体になんらかの影響を与えてしまうのが一般的で、仮に一般の生徒への余剰効果を口にしようとする、その配慮は教育の質を下げてしまうような類のものとしかいかえなくなるだろう。

2つめは、Differential Boost Hypothesis である (Fuchs & Fuchs 1999, 2001; Lewandowski et al. 2008; Sireci et al. 2005)。この仮説は、あるテストの受験や学習において、何らかの介入を行った場合、障害のある子・ない子いずれも得点の上昇が見られるが、その上昇率が障害のある子のほうが大きい場合(差分がある場合)は、この Differential Boost Hypothesis に該当する。この仮説は、次のような例をもって示される。

学習障害のある子どもとそうでない子どもに対して、算数のテストで、読み上げを実施した際の得点とそうでない時の得点の比較を行った。その結果、いずれの子どもたちも読み上げを実施したほうが得点が高くなった。また、障害のある子どもたちのほうが、大きく得点が上がった (Fuchs & Fuchs 1999)。

このように、いずれの子どもにとっても配慮の実施は利益があり、かつ障害のある子どもへの配慮効果が、障害のない子への配慮効果を上回っている場合は、その配慮が合理的に必要なで妥当であったことを立証していると説明

される。配慮を提供したことによって、障害のない学習者の得点が、これ以上上昇しなくなる「天井効果」をうまく制御し、障害のある子どもとない子どもとの間の差分を測定する必要があると Laitusis (2010) は指摘している。

ここで配慮として実施された読み上げによる、障害のある子どもとない子どもの得点の上昇という結果は、一見配慮によるもの(=「配慮効果」)のように見える。しかし、それは果たして本当に配慮による効果のみによるものといえるだろうか。例えば、学習障害の場合、これまで文字で学んでいた内容の理解が、文字の読み書きが困難であることで制限されていた可能性があり、その内容を音声、つまり読み上げで学んだ際に新しい学びを得たり、学習内容がその子どものなかで整理されていることは大いに考えられる。つまり、配慮として提供された読み上げは、配慮効果はもちろん、一定程度の教育効果を見込めるような「余剰効果」があるともいえるのではないだろうか。

もしも障害のない子どもも、その介入によって大きなメリットを得るとしたら、それはその受験者たちも、現在のテスト環境下で能力を最大限に発揮していなかったことになるという、妥当性の問題がこれまでに指摘されてきた (Kormos 2017)。また、その介入が、すべての障害のある子どもとない子どもの得点を向上させるわけではない (Lewandowski et al. 2007; Lin & Lin 2014; Händel et al. 2014)。その場合、すべての子どもたちに対してその介入を行うことは、公平性を十分に担保しているとはいえない。

3つめは Interaction Hypothesis (Sireci et al. 2005) である。この仮説は、ある配慮の実施によって、障害のある学習者のみが、配慮のない試験での得点と比較してより高い得点を出した場合が該当するものである。先に説明した通り、これまで test における妥当性と公平性、そしてその関係性は、様々に議論・定義されてきたが、特定の介入が公平な配慮であった場合、障害がある子どもと障害のない子どものグループの間と、配慮提供時と通常時の間になんらかの相互作用があるとの仮説は説明している (Cawthon et al. 2009; Sireci et al. 2005)。Maximum Potential Thesis は、包括的なテストではなく、試験時間の延長についてのみが言及されていたが、Interactive Hypothesis はその議論を引き継いだ仮説となっており、具体的な相互効果を制御することで、配慮による効果と、余剰とされる効果を区別しようという試みであるが、未だ道半ばといった趣が強いといえるかもしれない。

3-4. オンラインの時代における assessment

ここまで、assessment の3つの主要な理論を整理し、そのメリットとデ

メリットを比較してきた。本稿で特に強調したいのは、こと教育・学習の assessment の場合には、測定された結果のうちには2つの効果があり、その混濁が問題になってきた点である。通常、合理的配慮などの支援や介入が、それを受ける本人の能力そのものに何らかの影響を与えることはほとんどない。例えば、肢体障害における移動の評価を考えた場合、スロープ・エレベータなどの環境面の配慮や、電動車いすなどの技術支援があったとしても、それが該当者の歩行能力に影響するということはない。test の際にスマホやタブレットを使用したからといって、assessment では純粋に該当者の能力を評価することは、当然のごとく可能であるし、基本的な assessment は、本人の能力そのものが外的な影響を受けることがないことを前提として企画される。

しかし教育や学習に関して、評価や介入をしようとした場合は、そのような純粋な test や assessment を想定することができない。その test や assessment で評価された結果の中には、そもそも合理的配慮の影響とは別に、介入が実施されることで、該当者本人の能力そのものの向上（問題が分かりやすくなる、整理して理解しやすくなる、問題の構造把握が進み回答しやすくなるなど）という、「教育的な効果」の可能性があるからである。assessment に関する3理論は、全て、そのような、test や assessment からすると「余剰効果」とでも呼ぶべき効果を、実際の「合理的配慮の効果」と、どう切り離して評価するかという観点に傾注したものであった。そしてその営みは、全て報われたということではできず、現在でも公平性と妥当性の観点から、幾重にも異論を許しているといっよい。

確かに、test や assessment を、従来の他の能力評価と同質のものとして考え続ける限りは、この問題は常につきまとうであろう。しかし現代社会における教育の測定は、そのような見方を大幅に変更する必要性を現前させている。

一番の好例が、COVID-19 Crisis による、オンライン教育の急速な進展であろう。対面での授業を前提としていた場合と異なり、教育内容の全面的なオンライン化が必要となって、緊急で導入している教育機関は多数にのぼる。それらのオンライン化によって、障害学生・生徒らの学習環境も大きく変更を迫られている。デジタル化・オンライン化の特徴は、それが学習者全体に適用されるが、そのデバイス、アプリ、教材などの大半は、従来の対面授業よりも、多くの学生にとって環境改善に繋がっている。配布資料は拡大が自由で見やすくなり、音量は自分にマッチして自由に変更できたり、読み上

げの ON/OFF を選べたり、授業の内容は録音や録画できたり、オンデマンド教材を自由に見直したりできる。つまり、従来の拡大資料、音量変更、音声読み上げといった合理的配慮の一部が、自動的に全学生・生徒に配慮として提供される可能性が高く、しかもそれはさほど難しくないのである。このような状況では、assessmentにおいても、純粋な合理的配慮の成果と、オンラインが生む「余剰効果」とを区別することは、ほとんど不可能だろう。

つまりこの事態は、これまでの前述の3理論や、その母体となっている合理的配慮論が、どれほど対面での教育や試験を、無意識のうちに念頭においていたのかを、あらわしているということが出来る。急速にすすむオンライン化は、合理的配慮を受けた

教育評価の状況をも、大きく変容しているということが出来るのだ。それでは、そのような時代に有用な assessment 理論は、果たしてあり得るのだろうか。

4. まとめにかえて：オンライン化と assessment の時代に求められる評価理論としての Differential Boost Hypothesis

本稿では、まず、教育評価の前史を整理する中で、measurement や evaluation など集団内での比較素材として考えられてきたものが、それぞれの個別の成長を測定する assessment という方向に変容する傾向があったことを整理してきた。実は、test ではなく、assessment だけでよければ、合理的配慮策と混ざりあって、全体的な支援として「合理的配慮効果」と「余剰効果」が混濁しかねないオンライン化の assessment として独特の地位を得ることが出来る理論を見出すことができる。それが、Differential Boost Hypothesis である。

現在、主流となっている3理論のうち、Maximum Potential Thesis と Interaction Hypothesis は、「余剰効果」をいかに生み出さないか、というメソッドを目指している。確かに他者との比較という意味での公平性を厳密に確保しようとする場合、その戦略は間違っていない。しかし、各個人の成長を評価しようという assessment の場合は、そのような比較のための公平性は従属的な要素であり、むしろこだわることで、児童・生徒への教育全体の効果を制限しているとも考えることができる。子どもたちは、試験の中でも学び、試験をとおしても成長するはずなのだ。

一方で、Differential Boost Hypothesis は、「余剰効果」の発生を前提として、それを閾値まで制御し、残差から「合理的配慮効果」を見いだすという技法

であった。オンライン化の場合、このような「効果全体」を「教育効果」と考えることができれば、新しい時代の評価理論として、バージョンアップさせることができるのではないか。考えてみると、assessmentの結果に見られた成果は、環境構築による配慮の結果としても、「余剰効果」だとしても、それは一つの教育の総体がもたらした「教育の効果」といえるのである。配慮の有無ではなく、配慮の中で学び続けるのであれば、その環境形成も一つの効果とし、その後の生徒・児童の成長に目を向ければいいのではないか。つまりオンライン化の時代の assessment でよければ、環境配慮の結果も、余剰効果も、そして本人の努力の結果の成長も、全て教育の成果として把握することで、より合理的な成長の軌跡を描き出すことができるのである。

ここで示されているのは、配慮や環境の影響を含めた評価が切り拓きうる、新しい「能力」であるといえる。

その観点をどのように評価全体の中に位置付けるかは、今後の理論化と実践が待たれるが、それでも、環境や配慮を含めた能力全体を評価しうる Differential Boost Hypothesis がその土台となることは間違い無いだろう。古びた仮説として忘れ去られようとしている Differential Boost Hypothesis の再解釈が、個別具体的な assessment を求められながらも、急速なオンライン化・デジタル化の前で戸惑う私たちに、一筋の光を指し示してくれる。

謝辞

本研究は、科学研究費（基盤 C: JP20K12550）、津田塾大学私立大学ブランディング研究事業（社会的インクルージョン基盤研究）、公益財団法人日本生命財団・児童・少年の健全育成実践的研究助成の研究成果の一部である。ここで関連各位に深く感謝御礼申し上げる。

注

- 1) 心理学者 Terman に代表される知能の「遺伝説」は、その後それに対する「環境説」から、測定方法の非科学性や人種差別の特徴を指摘され、有名な論争になっている（1922年のリップマン＝ターマン論争）。
- 2) それまでの知能検査は、個別検査であった。
- 3) College Boardとは、米国の大学入試の標準テストの1つであるSATを主催する非営利団体である。
- 4) G・I法とは、1944年の復員兵援護法（Servicemen's Readjustment Act of 1944）の別名である。G・Iとは「アメリカ兵」を意味する俗語であり、この法律は、第二次大戦時アメリカの兵士だった人々

の、兵役後の権利や、保証などについて記載しており、この法律により中産階級へと社会階層を向上させたアフリカ系アメリカ人もいた (FDR Library 2020)。

- 5) 1971年、Pennsylvania Association for Retarded Children (PARC) が、ペンシルバニア州に対して起こした裁判である。当時ペンシルバニア州は、知的に障害のある子どもたちが無償の公立学校への入学を認めていなかった。これにより、公立教育を受けるのに知的に障害のある子どもたちが教育を受ける機会を猶予・制限したり、拒否するような政策は認められず、また6歳から21歳のペンシルバニア州の知的に障害のある子どもたちに教育があたえられるべきとされた (Abeson & Zettel 1977)。
- 6) 1972年、精神、行動、身体または情緒の障害を理由とし、公教育を与えられなかった子どもたちにかわり、保護者がワシントン D.C. 教育委員会を提訴した。このミルズ裁判の結果、すべての学齢期ににある子どもたちが、その障害の程度にかかわらず、無償で適切な公教育を受けることが認められた。ミルズ裁判は、知的障害児を射程としていた PARC 判決をすべての障害児に拡大したものであったと指摘されている (青木 2020)。
- 7) A Nation at Risk (『危機に立つ国家』) とは、1983年の連邦報告書である。この報告書には、当時のアメリカ全土の学力の現状と、それに対する提案が記され、アメリカの教育水準の低下 (危機) を警告している。SAT の得点が1960年代以降低下していることに加え、17歳のアメリカ人のうち、約13%が読み書きができないこと、特にマイノリティや貧困家庭にある子どもたちが学習上の困難を抱えていることなどが指摘された。
- 8) 厳密にいうと、Scholastic Aptitude Test から Scholastic Assessment Test へと名称変更したのは SAT I であり、従来 ACH (アチーブメントテスト) と呼ばれてきた教科別のテストは SAT II と名称変更し、この I と II のテスト2つを SAT としてまとめることにした (荒井 2003)。
- 9) 個別教育計画には、次のような内容も含まれる: 子どもに提供される特別な教育、関連サービス、補助的なエイドやサービス、プログラムの変更 (modification)、研修を含む、学校教職員に必要な支援など。

参考文献

- American Educational Research Association., American Psychological Association., National Council on Measurement in Education., & Joint Committee on Standards for Educational and Psychological Testing (U.S.), 1999, *Standards for educational and psychological testing*.
- American Educational Research Association., American Psychological Association., National Council on Measurement in Education., & Joint Committee on Standards for Educational and Psychological Testing (U.S.), 2004, *Standards for educational and psychological testing*.
- American Educational Research Association., American Psychological Association., National Council on Measurement in Education., & Joint Committee on Standards for Educational and Psychological Testing

- (U.S.), 2014, *Standards for educational and psychological testing*.
- American Psychological Association, 1954, Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin*, 51 (2, Pt.2), 1–38.
- American Psychological Association., American Educational Research Association., National Council on Measurement in Education., American Psychological Association., & American Educational Research Association, 1966, *Standards for educational and psychological tests and manuals*. Washington.
- American Psychological Association, 1974, *Standards for educational & psychological tests*. American Psychological Association.
- 青木亮佑 (2020)「アメリカにおける障害者教育法と障害者差別禁止法の関連性 —障害者教育における合理性修正をめぐる判例の検討を通じて—」法学研究論集, 52, 163-183.
- 荒井克弘 (2003)「学力評価システムの日米比較」教育社会学研究, 72, pp.37-52.
- 荒井克弘、倉元直樹 (2008)『全国学力調査 日米比較研究』金子書房 .
- Bachman, L. F., 1990, *Fundamental considerations in Language testing*, Oxford: Oxford University Press.
- Beaton, A.E., Rogers, A. M., Gonzalez, E., Hanly, M. B., Kolstad, A., Rust, K. F., Sikali, E., Stokes, L., and Jia, Y., 2011, *The NAEP Primer* (NCES 2011-463) . U.S. Department of Education, National Center for Education Statistics. Washington, DC.
- Camara, W., Copeland, T., & Rothchild, B., 1998, Effects of extended time on the SAT I: Reasoning Test: Score growth for students with learning disabilities. *College Board Research Report 98–7*. New York: The College Board.
- Cawthon, S.W., Ho, E., Patel, P.G., Potvin, D.C., & Trundt, K.M., 2009, Multiple constructs and effects of accommodations on accommodated test scores for students with disabilities. *Practical Assessment, Research, & Evaluation*, 14.
- Crouse, James; Trusheim, Dale, 1988, *The Case Against the SAT*. Chicago: The University of Chicago Press. pp. 16–39.
- ETS Committee for Testing Handicapped People, 1988, *Sourcebook for testing handicapped examinees*. Princeton, NJ: Educational Testing Service.
- FDR Library, 2020, Our Documents: The G. I. Bill, FDR Library. retrieved from <http://docs.fdrlibrary.marist.edu/odgist.html>
- Fischer, R. J., 1994, The Americans with Disabilities Act: Implications for measurement. *Educational Measurement: Issues and Practice*, 13, 17–26, 37.
- Fitzharris, L. H., 1993, *An Historical Review of the National Assessment of Educational Progress from 1963-1991*, Unpublished doctoral dissertation, University of South Carolina.
- Fuchs, D., & Fuchs, L. S., 1994, Inclusive schools movement and the radicalization of special education reform. *Exceptional Children*, 60 (4), 294–309.

- Fuchs, L. S. & Fuchs, D., 1999, Fair and unfair testing accommodations. *School Administrator*, 56, 24–29.
- Fuchs, L. S. & Fuchs, D., 2001, Helping teachers formulate sound test accommodation decisions for students with learning disabilities. *Learning Disabilities Research & Practice*, 16 (3), 174-181.
- 後上鐵夫、中澤恵江、松村勘由(2002)「Ⅲ 主要国における特別な教育的ニーズを有する子どもの指導について 5. アメリカにおける特別な教育的ニーズを有する子どもの指導に関する調査」『主要国の特別な教育的ニーズを有する子どもの指導に関する調査研究』国立特殊教育総合研究所.
- Henning, G., 1987, *A guide to language testing: Development, evaluation, research*. Boston, MA: Heinle and Heinle.
- Hollenbeck, K., Tindal, G., & Almond, P., 1998, Teachers' knowledge of accommodations as a validity issue in high-stakes testing. *Journal of Special Education*, 32, 175–183.
- Jennings, C. R., 2014, *Extended Time as a Testing Accommodation for Students with Disabilities*, Master's thesis, The University of Texas Southwestern Medical Center.
- 北野秋男(2015)「学力テストの暴力性 —アメリカにおける教育評価体制の歴史—」近代教育フォーラム, (24), 1-9.
- 北野秋男(2017)「現代米国のテスト政策と教育改革 —「研究動向」を中心に—」教育学研究, 84 (1), 27-37.
- Kubiszyn T. & Borich, G. D., 2015, *Educational Testing and Measurement*, 11th Edition, Wiley.
- Lewandowski, L. J., Lovett, B. J., & Rogers, C. L., 2008, Extended time as a testing accommodation for students with reading disabilities: Does a rising tide lift all ships? *Journal of Psychoeducational Assessment*, 26, 315-324.
- Lemann, M., 2004, “A History of Admissions Testing”. In Zwick, R. (ed.) , *Rethinking the SAT*, New York, Routledge.
- Lewandowski, L. J., Lovett, B. J., & Rogers, C., 2008, Extended time as a testing accommodation for students with reading disabilities: Does a rising tide lift all ships? *Journal of Psychoeducational Assessment*, 26 (4), 315-324.
- Lin, P.-Y., & Lin, Y.-C., 2014, Examining student factors in sources of setting accommodation DIF, *Educational and Psychological Measurement*, 74 (5), 759-794.
- Lovett, B. J., 2010, Extended time testing accommodations for students with disabilities: Answers to five fundamental questions. *Review of Educational Research*, 80 (4), 611-638.
- Messick, S., 1989, Validity. In R. Linn (Ed.) , *Educational measurement (3rd ed)*. Washington, DC: American Council on Education, 13-103.
- 村木英治(2003)「4つのテスト」『教育情報学研究』1, 109-113.
- Phillips, S. E., 1994, High-stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education*, 7, 93–120.

- Pitoniak, M. J., Royer, J. M., 2001, Testing Accommodations for Examinees with Disabilities: A Review of Psychometric, Legal, and Social Policy Issues, *Review of Educational Research*, 71, (1), pp.53-104.
- Ravitch, D., 2000, *Left Back: a Century of Failed School Reforms*, Simon & Schuster. (= 末藤美津子ら訳 (2008)『学校改革抗争の100年』東信堂)
- Rosenfield, M, Tannenbaum, R. F. J., Wesley, S., 1995, Policy Issues with Psychometric Implications in Impara, J, C. *Licensure Testing: Purposes, Procedures, and Practices*, Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska-Lincoln.
- Runyan, M.K., 1991, The effect of extra time on reading comprehension scores for university students with and without learning disabilities. *Journal of Learning Disabilities*, 24, 104-108.
- 渋谷憲一 (2003)『教育評価の基礎』教育出版
- Schudson, Michael S., 1972, "Organizing 'Meritocracy': A History of the College Entrance Examination Board," *Harvard Educational Review*, vol 42, No 1, pp.34-69.
- Shepard, L., Taylor, G., and Betebenner, D., 1998, Inclusion of limited-English-proficient students in Rhode Island's grade 4 mathematics performance assessment. CSE Technical Report 486, Los Angeles: University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.
- Sireci, S. G., Scarpati, S. E., & Li, S., 2005, Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research*, 75 (4), 457- 490.
- Stretch, L. S. & Osborne, J. W., 2005, Extended time test accommodation: Directions for future research and practice. *Practical Assessment, Research & Evaluation*, 10 (7), 1-8.
- 田中耕治 (1988)「測定・評価論—アメリカの教育測定運動の特徴 ～ターマンの足跡を中心に～」長尾十三二編『新教育運動の歴史的考察』明治図書、pp.116-33.
- 田中義郎 (2002)「SATの最新事情」『教育』257号、国土社、20-28.
- Terman, L. M., 1919, *The Intelligence of School Children*. Houghton Mifflin.
- Terman, L. M., 1923, *Intelligence Tests and School Reorganization*. World Book.
- Thorndike, E. L., 1918, The nature, purposes, and general methods of measurements of educational products, in G.M. Whipple, ed., *Seventeenth Yearbook of the National Society for the Study of education*, Public School Publishing, Bloomington, IL, 16-24.
- Tyack, David B., Cuban, Larry., 1995, *Tinkering Toward Utopia: A Century of Public School Reform*, Cambridge, Mass.: Harvard University Press.
- U. S. Department of Education, 2002, No Child Left Behind Act of 2001.
- U.S. Department of Education, 2019, Office of Special Education Programs, Individuals with Disabilities Education Act (IDEA) database.
- Weston, T.J., 2002, The validity of oral accommodation in testing. (NCES 200306). Washington, DC:

National Center for Education Statistics.

吉良直 (2009)「どの子も置き去りにしない (NCLB) 法に関する研究 —米国連邦教育法の制定背景と特殊性に着目して—」教育総合研究, 2, pp.55-71.

Zuriff, G.E., 2000, Extra examination time for students with learning disabilities: An examination of the maximum potential thesis. *Applied Measurement in Education*, 13 (1), 99-117.