大学入試における
英語外部テスト利用の再検討

鈴木 卓

Abstract:

Re-examination of the use of external English tests

in college entrance exams

In 2017, the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) announced its decision to use external English proficiency tests (e.g. TOEFL, EIKEN etc.) in the Common Test for University Admissions to be launched in AY 2020. This announcement triggered a fierce debate as to the appropriateness of such external English tests for use in the Common Test, leading to MEXT's decision to suspend the policy's implementation until 2024. This paper re-examines the six external English tests originally approved for use in the Common Test, using Bachman & Palmer's (1996) framework of test usefulness to allow for a systematic analysis. The results point to the usefulness of the external tests particularly in terms of construct validity and possibly of washback. However, they also reveal that there still remain issues to be resolved such as those of practicality and comparability of scores from different tests.

Key words

Common Test for University Admissions, test usefulness, external tests, communicative competence, four skills

要 旨：

2017 年に文科省は、2020 年開始予定の大学入学共通テストにおいて、TOEFL や英検等の外部の英語能力テストを利用する決定を発表した。この発表は、外部の英語テストを共通テストに利用することの適切性について激しい議論を引き起こし、結果としてこの方針は 2024 年までの実施見送りが決定された。本稿では共通テストへの利用が当初認められていた 6 種の外部テストについて、体系的な分析を行うためバックマンとパーマーのテスト有用性の枠組み(Bachman & Palmer, 1996)を用いて、再検討を行った。結果として、構成概念妥当性の面とおそらく波及効果の面

においても、外部テストの有用性が示された。しかし実際性や異なるテストの得点の比較可能性といった未だ解決すべき問題が残っていることも明らかになった。

キーワード：
大学入学共通テスト　テスト有用性　外部テスト　コミュニケーション能力
4 技能

1. はじめに

大学入試センター試験に代わり 2020 年度から実施予定の大学入学共通テスト（以下、「共通テスト」）の実施方針が 2017 年に発表され、民間の英語能力試験を利用する決定が公式に示された。従来も推薦入試等を中心に民間試験は利用されてきたが、大学入試センターが作成・実施する英語の試験に代わる「外部テスト」としての利用は新しい方針だった。文部科学省(2017)ではこの方針の決定理由を概略以下のように説明している。

・高校の指導要領で 4 技能育成が義務づけられており、大学入試でも 4 技能を評価すべきである。
・大学入試が 4 技能を問うものとなれば、高校の英語教育の改善促進が見込める。
・「話す」「書く」能力の試験の全国一律の実施は困難なため、民間試験の利用が必要である。

諸外国においても、全国規模の外国語試験は多肢選択式設問によって主に言語に関わる知識を問うものが従来は一般的だったが、近年では「話す」「書く」の言語産出技能を取り入れたテストの実施が増えている(Pižorn & Huhta, 2016)。4 技能を評価する外部テストの利用は、この意味で外国語教育の世界的な潮流に沿った方針転換だったと言える。

しかし 2019 年秋になってこの方針は実施見送りが発表された。発表された理由は、経済・地域的理由による受験機会不平等の是正に時間が必要だという一点だったが(萩生田, 2019)、外部テスト利用にはこの他にも様々な批判や懸念が示されている。今後の方針については一年間の再検討を行うとされ、その後外部テストは共通テストではなく各大学の個別入試で利用

させる方向で検討中との報道もある。このような状況においては、共通テストおよび個別入試における英語外部テスト利用の是非について、テスト理論の専門家だけでなく、入学者選抜の主体である各大学でも検討しておくべきだろう[1]。そこで本稿では、言語試験の有用性の概念 (Bachman & Palmer, 1996) を検討の枠組みとして用い、主として入学者選抜におけるテスト利用者としての視点から考察を行う。

言語テストの検証には様々な方法があるが、本稿で用いるバックマンとパーマーによるテスト有用性の枠組みは中でも主要な取り組みの一つである。近年ではバックマンらを含め、テスト検証は論証モデルを用いる方法に転換しつつあるが、テスト理論研究者以外にも利用しやすく、既成の複数のテストの利用検討という目的により合った前者の枠組みを本稿では用いる。

バックマンらは、ある言語テストが意図された目的を果たすために有効である度合いを「有用性 usefulness」という概念で表し、それは「信頼性 reliability」「構成概念妥当性 construct validity」「真正性 authenticity」「相互性 interactiveness」「影響 impact (また波及効果 washback)」「実際性 practicality」の 6 つの側面を持つとした。本稿ではこのうちテスト検証によく用いられる、信頼性、構成概念妥当性、影響（波及効果)、実際性の 4 要素を扱う。以下にまずバックマンらによる定義と例を引用しながら、各要素を簡単に説明する。なお本稿で言う「テスト」および同義の「試験」は、単に設問の集合にとどまらず、その開発・実施や採点・スコア解釈の過程をも含む。

信頼性：測定の安定性を意味する。例えば同種のテストの複数バージョンを同一受験者が受験した場合のスコアの安定性や、エッセイ等の自由回答式[2] 回答を複数の採点者が採点した場合のスコアの一致度を指す（後者の例はいわゆる「採点の揺れ」の有無に相当)。信頼性は一般に統計的手法で算出した信頼性係数（0 から 1 の間）で表現される。

---

[1] 外部テストはすでに多くの大学の入試制度において、推薦入試等の出願条件や選抜資料として利用されているが、本稿での「個別入試における外部テスト利用」とは、主に共通テスト利用入試や一般入試における利用を想定している。

[2] いわゆる「記述式」のことだが、口頭による回答も含むため本稿では「自由回答式」と呼ぶ。ただし、自由度は高くても通常は回答方法に一定の制約はある。

構成概念妥当性：テストのスコアを、測定したい言語能力（これを「構成概念 construct」と呼ぶ）の指標として解釈できる度合いを指す。例えばライティング科目の事前クラス分けテストにおいて、文法知識を問う選択式問題のみを出題すれば、文章を書くために必要な知識・技能の一部しか測定できず、構成概念妥当性は低くなる。

影響（波及効果）：テストは社会や教育制度に対して、また学習者や教員等の個人に対して、様々な「影響」を及ぼす。中でもテストが学習と指導の内容や方法等に及ぼす影響のことを「波及効果」と呼び[3]、そこには意図的なものと意図しないもの、また望ましいものや望ましくないものが含まれる。

実際性：人的・物的資源や時間等の面から、実施することが可能かどうか。人的資源とはテスト開発者・実施者・採点者等のことであり、物的資源とは場所・機材・資料等を指す。またこれらすべてに関して費用も考慮する必要がある。

このように検証角度を列挙し明確にすることによって、特定の文脈におけるテストの有用性の体系的検討・評価が可能となる。以下では、主にこのテスト有用性の枠組みを用い、外部テスト利用の是非や利用する場合の問題点について、批判的論点もふまえながら検討する。なおテスト開発者等が行う検証とは異なり、公表されているテストの仕様や統計データ等の資料を利用した間接的検証が主となることをお断りしておく。

2. 有用性理論から見た外部テスト

本項では、上のテスト有用性の枠組みに基づいて外部テスト（大別すると、IELTS アカデミック[4]、TOEFL、英検[5]、ケンブリッジ英検、GTEC、TEAP

[3] 影響と波及効果を区別せず同義で用いる立場もある。
[4] IELTS には他のバージョンもあるが、アカデミックのみが外部テストとして認定されたため、本稿では以下アカデミックを指して「IELTS」と呼ぶ。
[5] 例外的措置を除きコンピュータ使用の CBT 形式のみが外部テストとして認定されたため、本稿では以下 CBT 形式を指して「英検」と呼ぶ。

の 6 種）を検証し、問題点も検討しながらその利用の是非を検討する。

2.1 信頼性

日常的な「信頼」という言葉の意味とは若干異なり、テストの質における信頼性とは測定の安定性のことを指す。信頼性の検証には詳細な内部データが必要なため、ここでは公表された検証結果を拠り所として間接的に評価せざるを得ない。しかし幸いなことに主要な民間英語検定試験はこれまでも重要度の高い試験として使用されてきたため、試験の開発・実施を行う企業や団体（以下、「試験団体」）の中には、以前から信頼性検証や採点者の訓練を行い結果を公表しているものがある。おそらくもっとも早くからデータを公表してきた TOEFL を例にあげれば、信頼性検証において一般的な手法[6]を用いて検証を行い、総合点について信頼性係数 0.95 が得られたことを報告している(ETS, 2018)。なおセンター試験は信頼性係数を公表していないが、英語は 0.85 から 0.90 程度という推定や証言がある(Watanabe, 2013; 荘島, 2019)[7]。また TOEFL は自由回答式であるスピーキング部分の信頼性を 0.86、ライティングのそれを同 0.80 と報告している。言語テストに必要な係数の目安について「最低限必要なレベルは 0.8 だが、重要性の高いテストの場合には通常 0.9 以上が必要」(Weir, 2005, p. 29 筆者訳)とか、「信頼性 0.90 はとても良いとみなされる」(Chapelle, 2013, p. 4920 筆者訳)等とされることから、TOEFL 全体の信頼性は高く、また自由回答式部分だけを見ても許容範囲内の信頼性が確保されていると言える。IELTS、ケンブリッジ英検も同様に公式ウェブサイト上に信頼性係数を公表しており、IELTS 0.96、ケンブリッジ英検 0.92～0.95（級により異なる）[8]と十分な信頼性の確保を報告している。

これらに比較すると日本で開発されたテストは情報公開が限定的であり、GTEC、TEAP は総合点の信頼性係数は公表していない。ただし共に自由回答式セクションの信頼性を報告しており、GTEC はスピーキング部分の信頼

6 項目応答理論とクロンバック α 係数。総合点の係数算出法は不明。
7 ただしこれらの係数は問題数や受験者集団の特性等に影響を受けるため、テスト間の厳密な比較にはならない。
8 IELTS・ケンブリッジ英検共に合成信頼性（composite reliability）。

性係数が 0.95[9]、評価者間一致度が 79.4% (Koizumi, Okabe, & Kashimada, 2017)[10]、TEAP（パイロット版）はスピーキング部分・ライティング部分共に信頼性係数が約 0.95[11]、評価者間一致度が 57～60%（一段階の相違を無視すれば 97～100%）(Nakatsuhara, 2014; Weir, 2014)といずれも自由回答式部分として許容範囲内と言えるデータを公表している。GTEC、TEAP 共に「読む」「聞く」のセクションと総合点については信頼性係数を公表していない。英検は CBT 形式の信頼性データは示していないが、従来形式の試験について信頼性係数 0.82～0.9（級により異なる）という報告がある(Dunlea, 2015, p. 154)。文科省は外部テストの各試験団体に対し、採点の客観的検証と信頼性向上への改善努力を行い公表することを求めており(文部科学省, 2017)、また海外の試験団体はそれ以前からデータを公表している。日本の試験団体も自由回答式部分を中心にデータを公表し、実用上十分とされる信頼性の確保を報告しているが、テスト利用者としては総合点の信頼性係数を含め今後さらなる情報公開を望むところである。

このように大部分の試験団体が検証結果を公表しているにもかかわらず、外部テストの信頼性に対しては依然として懸念が示されている。その論点は主に二つあり、一つは外部テストが含む「話す」「書く」の設問や課題が自由回答式であることによる、採点の揺れへの懸念である。採点の揺れとは、一人の採点者が、同程度の能力を示す（と判断すべき）複数の回答に対して異なった評価をしたり、複数の採点者が、同一の回答を重複して採点した場合に評価が一致しないことを指す。採点の揺れはテストの信頼性低下の大きな要因となるため、それに対する懸念はまったく正当なものである。ただし同時に信頼性の一般的性質も認識しておく必要があり、一般に信頼性は、技能を発音・文法等のような要素に分割して個別に測定する場合に高くなりやすく、逆に分割せずに実際のコミュニケーションを模した形式で測定すると低くなりやすい(Davies, 1978/1982, Davies, 2014, p. 3 より引用)。つまりテストの信頼性を高くしようとすると、実際のコミュニケーションのあり方から乖離した設問・回答方式になりがちである。そもそ

9 受験者信頼性（person/test taker reliability）
10 執筆者の一名が試験団体所属。これ以降も同様に、試験団体所属等の執筆者が含まれる場合は、試験団体による調査として扱う。
11 受験者信頼性（person/test taker reliability）

も文科省が外部テスト利用を決めた理由に立ち戻れば、それは信頼性を追求し多肢選択式で「読む」「聞く」の能力を中心に測ってきた従来のセンター試験では、（次項で見るように）英語コミュニケーション能力が適切に評価できないことにあった。外部テストにせよ、あるいは大学入試センター等が新たに試験を開発するにせよ、自然の言語使用に近い形式で「話す」「書く」能力の評価を行えば、多肢選択式と比較して信頼性が低下することは避けられない。しかし上で見たように一定基準以上の信頼性が確保できるのであれば、それは入試に求める性質の優先順位に基づく選択と考えるべきだろう。

一方で外部テストの信頼性に関わる指摘には、その信頼性検証の作業が試験団体任せで、第三者による検証を経ていないという批判もある。確かに試験団体が自己検証したり外部研究者に検証を委託したりするだけでは、結果が偏向する可能性は否定できず、その検証プロセスについて「実施団体の公開情報を鵜呑みにしがたい」(羽藤, 2018, p. 60)という批判があるのももっともである。例えば自動車の燃費や排出ガスのデータの場合、メーカーによる調査・申告を基本としながらも国が抜き取り検査を行っており、その結果データの誤りや不正が発覚することもある。外部テストに関しても、特に共通テストの一部として用いる場合には、文科省が外部専門家に委託する等の方法で第三者による客観的検証を行うプロセスが必要ではないだろうか。もしそのような対応がされないまま各大学の個別入試に外部テスト使用が求められる場合には、大学側が各テストの自己検証結果や第三者による研究調査等を吟味して、信頼性の高い外部テストのみを選定することも検討すべきように思われる。

2.2 構成概念妥当性

構成概念妥当性とは、あるテストのスコアを、測定したい能力（「構成概念」と呼ぶ）の指標として解釈できる度合のことを指し、問題の内容や種類に加えて、外的変数（他のテストの得点等）との関連性、因子構造等の側面を含む。外部テストの試験団体は、信頼性と同様に自己検証を行い十分な妥当性の確保を報告しているが、問題内容面に関しては外部からも一定の検証が可能である。そこで概括的ではあるがここでその再検証を試み

る。そのためには、まず共通テストの英語試験がどんな英語能力を測るべきと意図されているか確認する必要がある。

文科省によると、共通テスト全体の目的は「高等学校における基礎的な学習の達成の程度」を判定しつつ「大学教育を受けるために必要な能力」についても把握することであり、知識・技能とともに「課題を解決するために必要な思考力・判断力・表現力等を評価する」(文部科学省, 2017, p. 2)。英語については「英語によるコミュニケーション能力の向上が課題となっており」と述べた上で、「読む」「聞く」「話す」「書く」の 4 技能を適切に評価できるようにする(同上, p.14)としている。

以上から、共通テストの英語で測るべき英語の能力は次のようにまとめられるだろう。

・言語の表現・媒介手段＝「読む」「聞く」「話す」「書く」の 4 技能
・一般認知的能力＝知識・技能に加え、思考力・判断力・表現力等
・言語能力の定義＝コミュニケーション能力
・出題範囲＝高校までの学習内容に加え、大学教育を受けるために必要となる能力

それぞれの側面について、外部テストの構成概念妥当性を手短に検討してみよう。まず言語の表現・媒介手段に関しては、導入の経緯から当然だが、6 種すべての外部テストが 4 技能の測定という条件を満たす。センター試験は設問の大部分が「読む」「聞く」の能力を測るものであり、発音や会話・作文の知識を問う問題も存在はしたが、多肢選択式という出題形式の制約上、実際に英語を話したり書いたりする技能ではなく、それに関する知識を間接的に評価せざるを得なかった。従ってセンター試験の英語は、英語コミュニケーション能力（後述）の評価手段としては、構造的に構成概念の代表性不足（出題される知識・技能等が構成概念全般から選ばれずに偏りがあること）の問題が避けられなかったと言える。外部テスト利用により「話す」「書く」能力のより直接的な測定を取り入れれば、表現・媒介手段に関して構成概念妥当性は大きく改善される。Messick(1996)は、リスニングやスピーキング問題のない試験は筆記試験が得意な受験者に有利であり、コミュニケーション能力のテストとして不公平であると指摘しているが、同様にセンター試験では「書く」「話す」技能が高い受験生は、そ

れをほぼ評価されないという不利益を被っていた。4 技能評価の外部テスト利用により、この点でより公平な評価が可能になる。

次に認知的能力の面では、「思考力・判断力・表現力等」を評価することが明示されている。このうち表現力については、「話す」「書く」技能のテストを自由回答式で行うことが、そのまま評価につながると言える。「思考力・判断力」については、外国語の分野では「暗記に頼らず、文脈に応じて内容を考えたり表現形式を組み立てたりする」というような能力を含むとすれば、やはり「話す」「書く」技能のテストがその測定に適したものと考えられる。さらに外部テストでは複数技能の統合的使用を必要とする課題によっても思考力・判断力の評価を実現しようとしており、例えば文章や図表を「読む」ことを通じて得た情報を統合・要約して「話す」「書く」ことを求める課題が出題されている。

ただし思考力・判断力の評価に関しては、注意すべき点が二点あるように思われる。まず、どの程度の思考力・判断力を英語試験の構成概念の一部とみなすかという問題である。従来から大学入試では、読解問題等で正解に到達するのに論理的思考が必要となる（いわゆる「行間を読む」）問題は出題されてきた。しかし例えば複雑な計算を要するような問題は、英語力以外の要素が介入するため一般には避けられてきた。共通テストの英語で思考力・判断力を問うと公表すれば、受験生もそのための準備をする必要が生じる上、外部テストによってその扱いが大きく異なる可能性もある。求める（測定する）「思考力・判断力」の程度を文科省や試験団体は明らかにすべきではないだろうか。

加えて、思考力等を問うために複合技能が必要な課題を出題する場合、4 技能を独立して評価できなければ、その意味で構成概念妥当性の低下を招く。例えば与えられた文章を読んでその情報に基づいて論述文を書く課題の場合、文章の難度によっては「書く」能力の評価が「読む」能力に依存してしまうことが起きうる。複合技能の作問を初期から行ってきた IELTS では、技能別スコアの独立性確保に配慮しているとするが(IELTS, 2019, p. 6)、今回調べた限りではその他の外部テストで同様の対策がとられているかはわからなかった。

次に言語能力の面では、共通テストの英語試験は「英語コミュニケーシ

ョン能力」を測定することを（少なくとも目標として）意図していると考えて良いだろう。コミュニケーション能力は、かつては文法知識等に対立するものとして、相手とやりとりを行う能力というように考えられたが、現代では「文法能力・談話能力・社会言語学的能力・方略的能力」から成る(Canale & Swain, 1980)というように構成要素を明確にして広く捉えることが一般的である。学習指導要領でもこれに似通った「語彙・文法などの知識と技能」「社会的文脈などを考慮して言葉を使える力」「場面・状況・相手などを考慮して言葉を使える力」から成るという CEFR の定義を引用していることから、これが共通テストで測るべき英語コミュニケーション能力と考えられる。

こういった意味でのコミュニケーション能力測定の実現という点では、同じ 4 技能試験でも外部テストは一様ではない。「話す」技能のテストを例にとると、試験形式が実生活でのコミュニケーション形態に比較的近いのは、面接官と直接またはオンラインでやりとりする IELTS やケンブリッジ英検であろう。一方でコンピュータの指示等に応じて受験者が回答を録音する TOEFL や英検、TEAP は、実生活とは類似性が低い状況で英語を話すことになる[12]。また面接官とやりとりを行う IELTS 等にしても、面接官はあいづち等を最小限に抑えるので[13]、現実のコミュニケーションと比較すると、相互作用が極めて少ない等の違いがある(Seedhouse & Egbert, 2006)。このような試験状況と実生活の言語使用の相違を重視して、「対人コミュニケーションは評価が難しく、数値では測定できない」(鳥飼, 2018)というような批判もある。しかし上述した短所があるとはいえ、外部テストの「話す」「書く」のテストは、談話構成能力や語用論的能力を（加えて「話す」場合には、流暢性や即興性も）問うという意味では現実のコミュニケーションの特性を反映している。少なくとも多肢選択式による間接的な「話す」「書く」能力の評価と比較すれば、外部テストの利用により構成概念妥当性が大きく向上することは間違いがない。

最後に、共通テストで問うべき出題範囲については、学習指導要領で定

12 とは言え情報通信技術(ICT)の進歩・普及により一般的なコミュニケーション形態も変化しており一概には言いにくい。
13 IELTS・ケンブリッジ英検両公式ウェブサイトで面接の動画が視聴できる。

める高校までの学習内容に加え、大学教育を受けるために必要となる能力を測るべきと定められている。本稿では学習指導要領と外部テストの内容の詳細な照合は難しいため、すでに整合性が確認されたという検証の方法を公表資料によって確認する。文科省(2018)は「（１）学習指導要領が育成を目指す能力と、各資格・検定試験において評価する能力に整合性があるか、（２）学習指導要領に基づく指導において取り上げられる言語使用の目的や場面と、各資格・検定試験が狙いとする言語使用の目的や場面に整合性があるか」の二点を中心に検証したとする。まず各試験団体がこの二点について自己点検を行った上で、その報告と実際の試験問題を基にして英語教育・高等学校英語教育課程それぞれの専門家と文科省の専門職員による整合性の確認がされた[14]。つまり学習指導要領との整合性の検証においては、文科省や外部専門家も作業に参加している点が、信頼性検証の場合とは異なっている。外部テストに対しては「民間英語能力試験に学習指導要領との整合性があるとは限らない」という批判があり、それじたいは正当な指摘だが、このような検証プロセスによってその可能性への対応がなされていることと、また当初から共通テストには学習指導要領の内容を超えて大学教育へのレディネスを問う内容も含むと企図されていることは認識しておくべきだろう。

ただしこのように学習指導要領との整合性は検証されていても、詳しく見るとそこには改善すべき点もある。それは上で見た文科省(2018)の（2）つまり「言語使用の目的や場面」に関連しており、各テストの目標言語使用領域が必ずしも明確に公表されていないことである。目標言語使用領域（target language use domain）とは、テスト外で受験者がその言語を使用するであろう状況や文脈を指し、テスト開発の際にはそれを考慮してテストに含める話題や課題を決めるべきとされる(Bachman & Palmer, 1996, p. 45)。目標言語使用領域は、広くは例えば「ビジネス・コミュニケーション」のように設定でき、より狭くは「オフィスでの勤務・管理」のようにも設定できる（同上)。外部テストの目標言語使用領域は「各資格・検定試験の実施概要」(文部科学省, 2019)等の文科省資料や、各テストのウェブサイトで、テストの「目的」「特長」等として説明されているが、両者の説明が明確で

[14] 検証作業の詳細結果が各試験団体のウェブサイトで公開されている。

なかったり必ずしも一致しない場合がある。以下に文科省(2019)と各試験団体ウェブサイトから特定できる目標言語使用領域を併記して比較してみる。（「概要」は文科省(2019)を、「HP」は各テスト公式ウェブサイトを指す）

・IELTS＝ 概要「英語の教育環境」、HP「英語圏での留学」[15]
https://jsaf-ieltsjapan.com/ielts/
・TOEFL＝ 概要「高等教育」、HP（英米加豪等におけると推定される）「大学で必要とされるアカデミックな英語」
https://www.ets.org/jp/toefl/test-takers/
・ケンブリッジ英検＝ 概要「実生活のさまざまな局面」、HP「実生活のさまざまな状況」
https://www.cambridgeenglish.org/jp/why-cambridge-english/
・GTEC＝ 概要は領域の特定なし、HP（日本の中高生のと推定される）「日常生活や学校、留学場面」https://www.benesse.co.jp/gtec/fs/question/
・TEAP＝ 概要「EFL 環境の大学で行われる授業等で行う言語活動」、HP「大学教育（留学も含む）」https://www.eiken.or.jp/teap/merit/index.html
・英検＝ 概要「英語圏における社会生活(日常・アカデミック・ビジネス) 」、HP「社会で求められる実用英語」「身の回りの日常会話から、教養を深める社会的な題材まで、実際に英語を使用する場面」
https://www.eiken.or.jp/eiken/merit/

文言の違いはやむを得ないとしても、目標言語使用領域がいわゆる英語圏に位置づけられているか否かという重要な点が、必ずしも明確にされていない。具体的には TOEFL と IELTS は、「概要」からは明らかではないが、使用領域を（日本のように）英語を外国語として使用する国に位置づけてはいない。また英検は逆に HP ではわからないが、概要では使用領域を「英語圏」に特定している。このように目標言語使用領域を特定する結果、例えば IELTS には次のような問題が含まれる（例は公式ウェブサイトのサンプル問題。筆者訳）

[15] HP では留学以外にも「就労・移住」のためとも述べられているが、就労・移住には別種の「IELTS ジェネラル」が対応している。

あなたは大学の寮をルームメイトとシェアしていますが、いろいろ問題がありなかなか勉強が進みません。大学の accommodation officer に手紙を書いて状況を説明した後、何が問題でなぜ勉強がしにくいのかを説明し、どのような住居を希望するか述べなさい。

日本の大部分の受験生にとって、このような状況は近似したものを含めても経験がないだろうし、手紙の相手である accommodation officer という職名になじみがなければ、コミュニケーション能力の一部である「語用論的能力」（場面・状況・相手などを考慮して言葉を使える力）を発揮できない。また「どのような住居を希望するか」と問われても、他にどんな選択肢があるのかがわからなければ答えることは難しい。このように、目標言語使用領域をいわゆる英語圏に位置づける試験の場合、日本の受験生の一般的言語使用状況にはない話題やタスクが含まれるため、それに向けた準備なしで受験すれば本来の英語力よりも評価が低くなりうる。後述するように目標言語使用領域がテスト間で異なること自体は、受験生にとってメリットもあるが、各テストの使用領域とその相違についてはより明確に受験生に周知する必要がある。

以上見たように、外部テストを共通テストの英語試験として用いる場合、全体としてセンター試験よりも構成概念妥当性は大きく向上すると考えられる。しかし学習指導要領との整合性については、検証を経てはいるものの、目標言語使用領域の多様性が受験生に不利に働く可能性があり、その点はさらに明確に周知することが必要であると考えられる。

2.3 影響と波及効果

テストは社会や教育制度および個人に対して様々な影響を与えることが知られている。本稿に関係深い例として、「国家規模で行われる重要な試験において、多肢選択式やある種の口頭面接といった特定の出題形式を用いることが、その国の教育方法や言語教育課程に与えうる影響」という例をバックマンらはあげている(1996, p. 34 筆者訳)。様々な影響の中でも「学習や指導」に関わる影響を波及効果と呼ぶ。波及効果や影響には意図的なも

のと意図しないもの、有益なものと有害なもの、結果に関するもの（能力の向上等）と過程に関するもの（学習方法の変化等）がある。Davies ら(1999, p. 225)は、有害な波及効果として「ライティング技能のテストを多肢選択式のみをもって行えば、ライティングそのものを練習するよりも、そのような設問へ答える練習をするように強い圧力が生じる」（筆者訳）という例をあげているが、これは非意図的で、過程に関する波及効果でもある。

波及効果についての初期の議論では、試験の種類や内容によって学習や指導の内容・方法が決まると断定的に語られる傾向があった。日本でも英語の入試問題に関しては、知識偏重の学習内容や丸暗記という学習方法を促し、英語運用能力の向上に結びつかないと批判されることが多かった。しかし 1990 年代からの実証的研究(Alderson & Hamp-Lyons, 1996; Watanabe, 1996 等)によって、波及効果は試験や学習・指導を取り巻く文化・文脈や、教員・学生の信念・属性等の要因によって変化する複雑な事象であることが示されている。このため現在では一般に波及効果はテスト使用の具体的な文脈に位置づけて検討する必要があると認識されている(Green, 2013, p. 49)。

ではここでセンター試験（英語）と外部テストそれぞれの波及効果についての研究を概観してみよう。センター試験の波及効果については、上述したように印象に基づいて語られることは多いが、実証的な研究は非常に少ない(Watanabe, 2013)。その大きな理由は比較対象（試験の影響を受けない集団）が見つけにくいことだと思われるが、センター試験にリスニングが導入された前後にその波及効果を調査した研究がある(Hirai, Fujita, Ito, & O'ki, 2013; 内田 & 大津, 2012; 斉田, 2013)。これらの研究によれば、リスニング導入により英語を聞くことへの学習意欲が増大したり、聞く力の伸びを学習者や高校英語教員が実感したりという有益な波及効果が認められた。しかし一方でリスニング導入前後の入学者を他のテストを用いて比較してみると、リスニング能力に有意差はなく（またはわずかで）、客観的に見た「結果」面での効果は明らかではないことが報告されている。

外部テストとして認定された試験の波及効果を調べた研究としては、TOEFL に関しての Alderson & Hamp-Lyons(1996)、Wang(2019)や、IELTS に関しての Allen(2016, 2017)、Green(2007)、Hawkey(2006)、TEAP についての

Green(2014)、Sato(2018)、英検 2 級（従来形式）に関しての藤田他(2016)等がある。このうち日本人学習者を対象に行われた研究においては、やはり学習動機や自己評価における好ましい影響が報告されているが、「話す」「書く」能力に向上が見られたという報告は Allen (2016, 2017)による「話す」能力の向上に限られ、ここでも客観的な「結果」面の変化を示すエビデンスは多くはない。また指導・学習内容や方法の変化については、言語産出技能の学習が増えたという Allen による報告(同上)もあれば、授業や学習の内容と試験の因果関係に対してより慎重な見解もある(Green, 2014; Sato, 2019)。

上述したように、波及効果はテスト単体の属性ではなく特定の文脈に位置づけて検討する必要があるため、先行研究から外部テスト利用の波及効果を予測することは容易ではない。また波及効果に関する研究一般に対して、その多くが小規模な調査であることや、大規模な調査は試験団体主導のものであること、また対照群が設定されていない研究が多いこと、等に対して方法論上の疑問も提出されている(寺沢, 2019)。とは言え、これらの研究から得られた知見を総合的に評価すれば、外部テスト利用によって学習者の学習動機や自己評価は好ましい影響を受ける可能性が高く、学習・指導の内容や方法もその可能性があるが、客観的に測定した「話す」「書く」能力の向上に直結するかどうかは確かではない、と考えて良いだろう。外部テスト導入についての議論の中で、それが「話す」「書く」能力の向上につながると当然のように主張されたり議論の前提とされることがあるが、それを無批判に受け入れるべきではないということになる。しかし逆に外部テスト利用によって、高校での英語の授業が試験対策に追われ、本来あるべき姿から離れるというような主張(鳥飼, 2018)も、その根拠が明らかではないように思われる。

2.4 実際性

実際性とは、人的・物的資源や時間・コスト等の面から、特定のテストを実施することが可能であるかどうかを指す。外部テストの大部分はこれまで民間英語能力試験として実施されてきたから、ここで問題となりうるのは、それらを共通テストまたは各大学の個別試験として適切に実施可能かということである。外部テスト利用についての議論の中で、実際性につ

いて論点となるのは主に受験機会の不公平性についてであり、これが外部テスト利用見送りの直接の要因となったことから、ここではその点についてのみ取り上げる。

外部テストの受験機会については、地域・経済的不公平性の問題が主に指摘されてきた。地域的不公平性とは、外部テストの受験会場が都市部に偏っており、山間部や島嶼部に居住する受験生には受験のための経済・時間的負担が大きいことを指す。経済的不公平性とは、外部テストの受験料がセンター試験受験料に加えて必要となるため、経済的事情により（練習のための受験も含め）受験機会が不平等になるということを指す。もとよりこの二つの問題は、外部テストの利用に関わらずセンター試験単独についても程度の差こそあれ存在していた。センター試験の会場も都市部に偏る傾向があり、また予備校等の模擬試験や対策講座も経済的事情により利用できる機会は平等ではないため、外部テストの利用によってこれらの問題が新たに生じるように捉えるべきではない[16]。とは言え、外部テストの受験が共通テストの他科目に加えて必要となることにより、この不公平性がいっそう増幅されるのは間違いがない。文科省もこの問題は認識し、離島・へき地の居住・通学者や経済的に困難な者等には受験期間の前倒しを認めたり、試験団体に対して低所得者世帯の受検者の検定料減免を要請する等したが、最終的には文科相が「経済的な状況や居住している地域にかかわらず、等しく安心して受けられるようにするためには、更なる時間が必要」(萩生田, 2019)と述べ、この受験機会の不公平性こそが外部テスト利用見送りの決定的理由となった。

上で見たように受験機会の不公平は大きく二つの要因から生じるが、地域的不公平についても、交通費や滞在費等の援助を行ったり、実施費用の公的援助によってへき地でも外部テストを実施したりすれば、受験生の負担は軽減できる。したがって地域的・経済的不公平性については、いずれも公的な費用負担によって改善が可能であるだろう。高等学校や高等教育の一部無償化が開始されたことを考慮すれば、外部テストの受験費用を地域・経済的状況に応じて一定額まで支給することは、特段に実現の難しい

16 東京都小笠原諸島の受験生は、センター試験受験のために約 3 週間内地に宿泊する必要があったという（石渡，2016）。

対応ではないように思われる。特に、共通テストに外部テストを組み込む場合に、そのような措置が早急に具体化されるべきであると考える。

3. 評価互換性

外部テスト利用に対する重要な批判的論点の一つに、複数の外部テスト間の評価互換性の問題がある。評価互換性は、テスト有用性の枠組みにおける構成概念妥当性の一側面（外的変数との関連性）と捉えることも可能だが、本来それが指すものとはやや異質なためここで項目を改めて検討する。外部テストには大別して 6 種の民間英語能力試験が認定されているが、それぞれ満点や難易度等が異なるため、そのままでは別種の試験のスコアを比較できない。そこで文科省では評価の統一基準として CEFR（ヨーロッパ言語共通参照枠）を指定し、各テストの特定のスコアや級が CEFR のどの段階に該当するかという対照表を作成した。

しかしこの対照表については強い批判がある。大別すると、CEFR 自体およびそれを入試目的に使用することへの批判と、CEFR と各テストとの対応の調査方法の問題に分けられる。CEFR は、ヨーロッパにおける外国語学習や教授・評価のための共通の参照枠として作成されたものであり、特定言語の能力評価のために開発されたものではない。また外国語能力のレベルを示すために、一般に Can Do ステートメントと呼ぶ質的評価指標を用い、初級者から母語話者レベルまでの習熟度を 6 段階に分けて記述する（現在ではより細分化する場合もある）。各段階の指標となる Can Do ステートメントは、例えば次のようなものである。

> Can exchange, check and confirm information, deal with less routine situations and explain why something is a problem. 情報を交換・確認・裏付けることができ、型通りではない状況にも対応でき、何かがなぜ問題なのかを説明できる（筆者訳）

このような CEFR に対しては、Can Do ステートメントをどの程度達成すればその段階に達したと言えるのか不明なことや、必ずしも現実の言語獲得の順序と合致しないことが指摘されてきた(Fulcher & Davidson, 2007, p. 100)。このため大学入試への利用についても、CEFR は「特に学習者自身の自己評価を重視している」、「厳密どころか相当に緩やかな尺度である」(鳥

飼, 2018)という指摘がある。またこれを日本の大学入試に利用した場合、ほとんどの受験者は 6 段階評価のうち下の 3 段階（A1～B1 と呼ばれ英検 3 級から 2 級程度に相当）に含まれるので、同評価となる受験者が多数となり、入学者選抜の目的には有効ではないという指摘もある(南風原, 2018)。

また CEFR と各テストのスコア・級等の対応付けの根拠についても批判がある。文科省では各テストのスコア等と CEFR 各段階の対応関係を示した対照表を作成したが、そのために必要な対応関係の調査作業が外部テストの試験団体に任されており、そこに文科省や第三者による検証が実質的にはなかったことが問題視されている。各試験団体のウェブサイト等で公開されている対応付けの方法を参照すると、一般的な手続きに従ってはいるものの作業者や作業方法が統一されていないため、各テストで CEFR の解釈が異なったり、対応付けの方法に揺れがあったりした可能性は高い。さらには対照表じたいが短期間に何度も更新されたこともあって、その作業の正確さにも疑問が持たれている(羽藤, 2018 等)。また構成概念妥当性の項で述べたように、外部テストは目標言語使用領域が一様ではないことから、理論的にも同一受験者の得点がテストの種類により異なることが予想される。

ただしこのうち、外部テストの種類が多く言語使用領域が異なることにはメリットもある。それは例えば英語圏への留学希望者は TOEFL 等を受験し、国内の国際系学部等への進学希望者は TEAP を選ぶというように、学習者が自らの目的や経験に応じて特定の分野に重点を置いた（またはより汎用的な）試験を選択できることである。受験準備に多大な労力を費やす大学入試においてテストの自発的な選択が可能となれば、学習動機や英語力のいっそうの向上が期待でき、これは入学者選抜だけを目的としない外部テストを利用する際のメリットと言える。また同一教科において複数の試験の選択肢があることは、センター試験でも社会科や理科、外国語等において同様であり、その場合は内容や平均点が異なる試験が同一教科の得点として用いられてきた[17]。このようなことを考慮すると、外部テストの難易度や内容に相違があり、テスト間の得点の対応が厳密なものではなくて

[17] ただし大学や学部によって受験科目が指定される場合があり、また科目間に著しい平均点の差がある場合は、得点が調整されることがまれにあった。

も、受験生がそれを十分に理解しており、また選択の自由が確保されている限りは公平性は担保されており、むしろそのメリットを生かすべきであると考える。

ただここで CEFR の入試利用や現行の CEFR 対照表についての批判に立ち返ってみれば、その論点はいずれも妥当なものである。少なくとも現行の対照表を 6 段階評価の形のまま入試に利用しても、受験者を序列付ける弁別力があるかどうかは疑わしく、評価段階の細分化等の対応が必要となるだろう。また各テストの評価互換のための仕組みも、さらに時間をかけても再検討されるべきであろう。試験団体の垣根を越え、第三者の専門家の手により、作業方法を統一して CEFR との対応を再検討したり、2 つのテスト間でしばしば行われているスコア相関性の調査を拡大して行う等の方法で、対照表をより精密かつ細分化したものにすることは可能でありまた必要であると考える。その上で、多数のテスト間の対応に一定の誤差や齟齬が残ることは避けられないとしても、受験生がテストを自由に選択可能であれば公平性は担保でき、また選択によって生じる英語学習上の利点もある。したがって解決すべき課題は、各テストの言語使用領域の相違やテスト間のスコア対応についてより精密・明確化した上で情報公開を徹底することと、受験料の公的補助等によってテスト選択の自由を保証することであろう。各大学の個別入試に外部テストを利用する場合にも前者は必要であるだろうし、共通テストの英語試験として利用する場合には後者も含めて実施することが強く望まれる。

4. まとめ

本稿では外部テスト利用の是非について、テスト有用性の枠組みに沿って、テスト利用者の立場から再検討を行った。4 技能を測る外部テスト利用の大きな利点としては、構成概念妥当性が向上することと、それにより、従来不利益を被っていた「話す」「書く」能力の高い受験生が正当に評価されることがあげられる。波及効果の面でも、高校生が英語で話すことや書くことの学習に積極的に取り組んだり、そうすることに自信を持つという効果が期待できる。さらに複数の外部テストを利用可能にすることによって、受験生が目的に応じてテストを選択できることもメリットである。一

方で、高校の授業の方法や内容が大きく変わったり、学習者の「話す」「書く」力が著しく向上したりといった効果を得るためには、外部テスト利用という改革だけで十分かどうか定かではない。また、特に外部テストを共通テストの一部として利用する場合には、受験機会の不公平性やテスト間の評価互換性の点について、文科省や試験団体によるさらなる対応が必要である。大学の個別入試に外部テストを利用する場合でも、評価互換性や各テストの信頼性については、選抜主体である各大学においてさらなる検証を行うことが望ましいであろう。また本稿で試みたように外部テストを検証してそこに高い品質を求めるのは当然のことだが、もし従来のように大学個別の入試問題作成を継続する場合には、外部テスト同様に採点基準の検討を含めた信頼性の向上や妥当性の検証を行い、各大学においても入試の有用性を高めるためのいっそうの努力をすべきであろう。

引用文献

石渡嶺司 (2016) 「センター試験で 24 泊 25 日！～離島高校生の受験格差を考える」Yahoo JAPAN ニュース https://news.yahoo.co.jp/byline/ishiwatarireiji/20160123-00053703/

内田照久、大津起夫 (2012) 「大学入試センター試験への英語リスニングテストの導入に至る歴史的経緯とその評価」『日本テスト学会誌 9(1)』77-84

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接続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』

荘島宏二郎 (2019) 「センター試験「英語」はどのような試験だったか」『大学入学者選抜における英語試験のあり方をめぐって(2)シンポジウム報告書』東京大学

寺沢拓敬 (2019) 「これからの英語教育の話を続けよう第 15 回「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」論議」『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/

鳥飼玖美子 (2018) 『英語教育の危機』筑摩書房（電子図書 Kindle 版）

南風原朝和 (2018) 「英語入試改革の現状と共通テストのゆくえ」南風原朝和(編)『検証 迷走する英語入試―スピーキング導入と民間委託』(pp.

5−25). 岩波書店
萩生田光一 (2019) 「大臣メッセージ（英語民間試験について）」（報道資料）
https://www.mext.go.jp/content/1422381_01.pdf
羽藤由美 (2018) 「民間試験の何が問題なのか―CEFR 対照表と試験選定の検証より」南風原朝和(編)『検証 迷走する英語入試―スピーキング導入と民間委託』 (pp. 41-68) 岩波書店
藤田亮子、横内裕一郎、松岡大地、仲村圭太、平井明代 (2016)「英検 2 級のテスト問題の分析−CEFR レベル、学習到達目標、波及効果の観点から」『関東甲信越英語教育学会誌, 30』85-97
文部科学省 (2017) 「大学入学共通テスト実施方針策定に当たっての考え方」
https://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2017/10/24/1397731_002.pdf
文部科学省 (2018) 「高等学校学習指導要領と英語資格・検定試験との関係について」
https://www.mext.go.jp/content/20200318-mxt_daigakuc02-000005103_5.pdf
文部科学省 (2019) 「各資格・検定試験の実施概要」
https://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2019/09/05/1420230_2_1_1.pdf

Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: a study of washback. *Language Testing, 13*(3), 280-297.
Allen, D. (2016). Investigating washback to the learner from the IELTS test in the Japanese tertiary context. *Language Testing in Asia, 6*, 1-20.
Allen, D. (2017). *Investigating Japanese undergraduates' English language proficiency with IELTS: Predicting factors and washback*. IELTS Research Partnership Papers, 1-56.
Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.
Canale, M., & Swain, M. (1980). Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing. *Applied Linguistics, 1*(1), 1-47.
Chapelle, C. A. (2013). Reliability in Language Assessment. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 4918-4923). Oxford: Blackwell/Wiley.
Davies, A. (2014). Fifty Years of Language Assessment. In A. J. Kunnan (Ed.), *The Companion to Language Assessment* (pp. 1-21) Chicester : John Wiley & Sons.
Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of Language Testing.* Cambridge: Cambridge University Press.
Dunlea, J. (2015). *Validating a set of Japanese EFL proficiency tests :*

demonstrating locally designed tests meet international standards. (PhD Thesis). University of Bedfordshire.

ETS. (2018). *Reliability and Comparability of TOEFL iBT® Scores*. https://www.ets.org/s/toefl/pdf/toefl_ibt_research_s1v3.pdf

Fulcher, G., & Davidson, F. (2007). *Language Testing and Assessment: An Advanced Resource Book*. London and New York: Routledge.

Green, A. (2007). *IELTS washback in context: preparation for academic writing in higher education.* Cambridge: Cambridge University Press.

Green, A. (2013). Washback in language assessment. *International Journal of English Studies, 13*(2), 39-51.

Green, A. (2014). *The Test of English for Academic Purposes (TEAP) impact study: report 1 - preliminary questionnaires to Japanese high school students and teachers*. Eiken Foundation of Japan. http://hdl.handle.net/10547/338488

Hawkey, R. (2006). *Impact theory and practice : studies of the IELTS test and Progetto Lingue 2000*. Cambridge: Cambridge University.

Hirai, A., Fujita, R., Ito, M., & O'ki, T. (2013). Washback of the Center Listening Test on Learner's Listening Skills and Attitudes. *ARELE: Annual Review of English Language Education in Japan, 24*, 31-45.

IELTS. (2019). *Guide for teachers: Test format, scoring and preparing students for the test.* https://www.ielts.org/-/media/publications/guide-for-teachers/ielts-guide-for-teachers-uk.ashx

Koizumi, R., Okabe, Y., & Kashimada, Y. (2017). A Multifaceted Rasch Analysis of Rater Reliability of the Speaking Section of the GTEC CBT. *ARELE: Annual Review of English Language Education in Japan, 28*, 241-256.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241-256.

Nakatsuhara, F. (2014). *A Research Report on the Development of the Test of English for Academic Purposes (TEAP) Speaking Test for Japanese University Entrants - Study1 & Study2*. Eiken Foundation of Japan. https://www.eiken.or.jp/teap/group/pdf/teap_speaking_report1.pdf

Pižorn, K., & Huhta, A. (2016). Assessment in educational settings. In D. Tsagari & J. Banerjee (Eds.), *Handbook of Second Language Assessment* (pp. 239-254). Berlin: De Gruyter.

Sato, T. (2018). The Impact of the Test of English for Academic Purposes (TEAP) on Japanese Students' English Learning. *JACET Journal, 62*, 89-107.

Sato, T. (2019). An investigation of factors involved in Japanese students' English learning behavior during test preparation. *Papers in Language Testing and Assessment, 8*(1), 69-95. http://www.altaanz.org/uploads/5/9/0/8/5908292/8_1_s4_sato.pdf

Seedhouse, P., & Egbert, M. (2006). The interactional organisation of the IELTS Speaking Test. *IELTS Research Reports Volume 6.* IELTS Australia and British Council.

https://www.ielts.org/-/media/research-reports/ielts_rr_volume06_report6.ashx
Wang, Y. (2019). The Impact of TOEFL on Instructors' Course Content and Teaching Methods. *The Electronic Journal for English as a Second Language, 23*(3), 1-18.
Watanabe, Y. (1996). Investigating washback in Japanese EFL classrooms. *Australian Review of Applied Linguistics. Series S, 13*, 208-239.
Watanabe, Y. (2013). The National Center Test for University Admissions. *Language Testing, 30*(4), 565-573.
Weir, C. J. (2005). *Language Testing and Validation: An Evidence-based Approach*. Hampshire: Palgrave Macmillan.
Weir, C. J. (2014). *A Research Report on the Development of the Test of English for Academic Purposes (TEAP) Writing Test for Japanese University Entrants*. Eiken Foundation of Japan. https://www.eiken.or.jp/teap/group/pdf/teap_writing_report.pdf

ウェブサイトの最終閲覧日はすべて 2020 年 5 月 5 日